# Achievements & Challenges of Deep Learning

## *from speech to language/multimodality*

**Li Deng**

*Deep Learning Technology Center, Microsoft Research, Redmond, WA. USA*

*Keynote, Sept 18, 2014 at Interspeech*

**Acknowledgements:** Geoff Hinton, Dong Yu, Xiaodong He, Jianfeng Gao, Yelong Shen, Xinying Song, Jinyu Li, Yi-fan Gong, Frank Seide, Mike Seltzer, Qiang Huo, Alex Acero, George Dahl, A. Mohamed, Po-sen Huang, Vincent Vanhoucke, Andrew Senior, Tara Sainath, Brian Kingsbury, John Bridle, Nelson Morgan, Hynek Hermansky, Paul Smolensky, Chris Manning, Eric Xing, Chin-Hui Lee, John Hershey, Mari Ostendorf, Les Atlas

# Main message of this talk

**Deep Learning**

**≈**

**Neural Networks, Deep**

in space & time (recurrent)

**+**

**Generative Models, Deep**

in space & time (dynamic)

**+**

…, …, …

# Main message of this talk

**Deep Learning**

**(converging with probability one)**

$\mathcal{F}$ **[Deep-Neural-Network, Deep-Generative-Model]**

**Two apparently opposing approaches: with highly complementary strengths and weaknesses**

| | Deep Neural Nets | Deep Generative Models |
|---|---|---|
| **Structure** | Graphical; info flow: **bottom-up** | Graphical; info flow: **top-down** |
| **Incorp constraints & domain knowledge** | Hard | **Easy** |
| **Semi/unsupervised** | Harder | Easier |
| **Interpretation** | Harder | **Easy** (generative "story") |
| **Representation** | **Distributed** | Localist (mostly) |
| **Inference/decode** | Easy | Harder (but note recent progress) |
| **Scalability/compute** | **Easier (regular computes/GPU)** | Harder (but note recent progress) |
| **Incorp. uncertainty** | Hard | **Easy** |
| **Empirical goal** | Classification, feature learning, … | Classification (via Bayes rule), latent variable inference… |
| **Terminology** | Neurons, activation/gate functions, weights … | Random vars, stochastic "neurons", potential function, parameters … |
| **Learning algorithm** | A single, unchallenged, algorithm -- BackProp | A major focus of open research, many algorithms, & more to come |
| **Evaluation** | On a black-box score – end performance | On almost every intermediate quantity |
| **Implementation** | Many untold-tricks | More or less standardized |
| **Experiments** | Massive, real data | Modest, often simulated data |

J. Acoust. Soc. Am. 121 (2), February 2007

# Speech production knowledge in automatic speech recognition

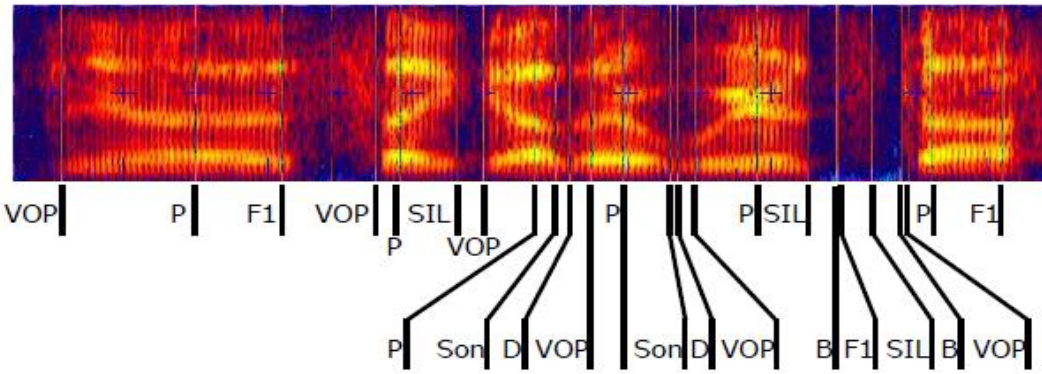
Simon King[a] and Joe Frankel
*Centre for Speech Technology Research, University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW, United Kingdom*

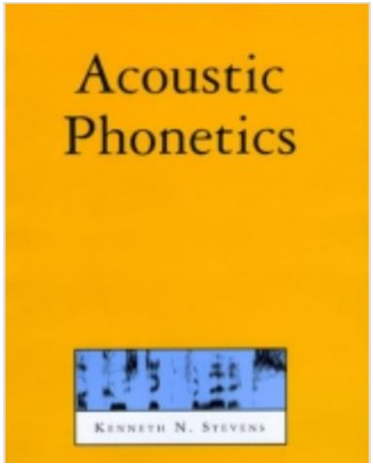
Karen Livescu
*MIT Computer Science and Artificial Intelligence Laboratory, 32 Vassar Street, Room 32-G482, Cambridge, Massachusetts 02139*

Erik McDermott
*Nippon Telegraph and Telephone Corporation, NTT Communication Science Laboratories, 2–4 Hikari-dai, Seika-cho, Soraku-gun Kyoto-fu 619-0237, Japan*

Korin Richmond and Mirjam Wester
*Centre for Speech Technology Research, University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW, United Kingdom*

(Received 13 October 2005; revised 1 November 2006; accepted 6 November 2006)




## MOVING BEYOND THE 'BEADS-ON-A-STRING' MODEL OF SPEECH

ASRU-1999

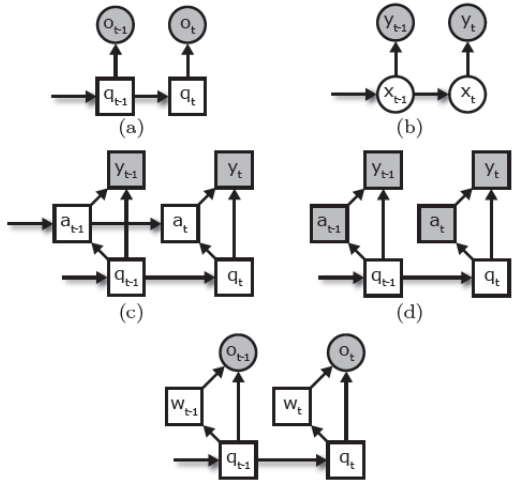
*M. Ostendorf*

Department of Electrical Engineering
University of Washington, Seattle, WA 98195

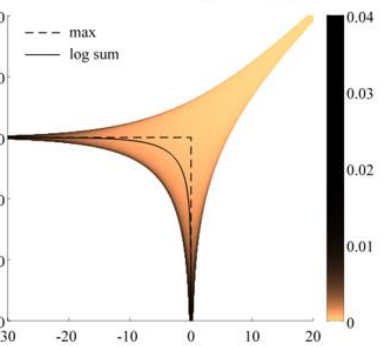
Computational Models of Speech Pattern Processing
NATO ASI Series Volume 169, 1999, pp 199-213

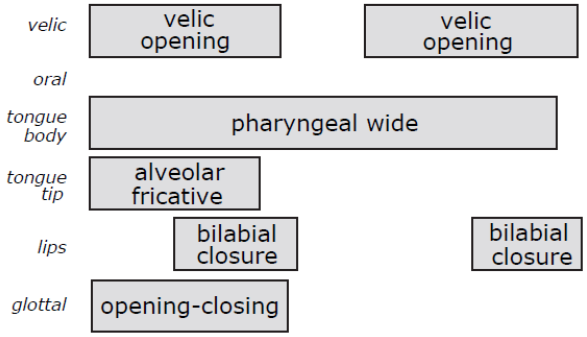
## Computational Models for Speech Production

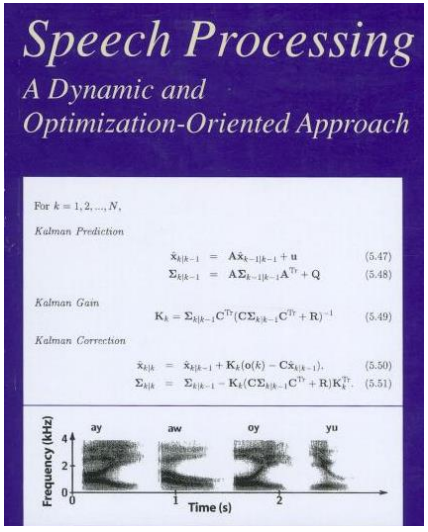
$$\frac{d}{dt}\begin{pmatrix}\mathbf{z}(t)\\ \dot{\mathbf{z}}(t)\end{pmatrix}=\begin{pmatrix}0 & 1\\ -\mathbf{S}^2(t) & -2\mathbf{S}(t)\end{pmatrix}\begin{pmatrix}\mathbf{z}(t)\\ \dot{\mathbf{z}}(t)\end{pmatrix}+\begin{pmatrix}0\\ \mathbf{S}^2(t)\mathbf{Z}^0(t)\end{pmatrix}+\begin{pmatrix}0\\ \mathbf{w}(t)\end{pmatrix}$$




$$p_y(\mathbf{y}|\mathbf{x},\mathbf{n},\mathbf{h})=\frac{1}{2}\left|\mathrm{diag}\left(e^{\mathbf{y}-(\mathbf{n}+\mathbf{x}+\mathbf{h})/2}\right)\right|$$
$$\mathcal{N}\left[\frac{1}{2}(e^{\mathbf{y}-(\mathbf{n}+\mathbf{x}+\mathbf{h})/2}-e^{(\mathbf{n}-\mathbf{x}-\mathbf{h})/2}-e^{-(\mathbf{n}-\mathbf{x}-\mathbf{h})/2});\mathbf{0},\boldsymbol{\Sigma}_\alpha\right]$$



# Outline

- **Part I:** A (brief) early history of 'deep' speech recognition
- **Part II:** Deep learning achievements: **speech** & vision
- **Part III:** Deep learning challenges: **Natural language**, **multimodality**, reasoning, mind, & deep intelligence in the big-data world

# Outline

- **Part I:** A (brief) early history of 'deep' speech recognition
  - (shallow) neural nets in ASR, before 2009
  - (deep) generative models in ASR and ML, before 2009
  - how/why DNN made recent inroad into ASR
  - roles of academic-industrial collaboration

- **Part II:** Deep learning achievements: **speech** & vision

- **Part III:** Deep learning challenges: **Natural language**, **multimodality**, reasoning, mind, deep intelligence in the big-data world

# Neural Networks in ASR
## (prior to the rising of deep learning)

**Temporal & Time-Delay (1-D Convolutional) Neural Nets**

- Atlas, Homma, and Marks, "An Artificial Neural Network for Spatio-Temporal Bipolar Patterns, Application to Phoneme Classification," NIPS 1987. — 1987
- Waibel, Hanazawa, Hinton, Shikano, Lang. "Phoneme recognition using time-delay neural networks." IEEE Transactions on Acoustics, Speech and Signal Processing, 1989. — 1989

**Recurrent Neural Nets** — 1991

- Bengio. "Artificial Neural Networks and their Application to Speech/Sequence Recognition", Ph.D. thesis, 1991.
- Robinson. "A real-time recurrent error propagation network word recognition system," ICASSP 1992. — 1992

**Hybrid Neural Nets-HMM**

- Morgan, Bourlard, Renals, Cohen, Franco. "Hybrid neural network/hidden Markov model systems for continuous speech recognition," IJPRAI, 1993. — 1993

**Neural-Net Nonlinear Prediction**

- Deng, Hassanein, Elmasry. "Analysis of correlation structure for a neural predictive model with applications to speech recognition," *Neural Networks*, vol. 7, No. 2, 1994. — 1994

**Bidirectional Recurrent Neural Nets**

- Schuster, Paliwal. "Bidirectional recurrent neural networks," IEEE Trans. Signal Processing, 1997. — 1997

**Neural-Net TANDEM**

- Hermansky, Ellis, Sharma. "Tandem connectionist feature extraction for conventional HMM systems." ICASSP 2000. — 2000
- Morgan, Zhu, Stolcke, Sonmez, Sivadas, Shinozaki, Ostendorf, Jain, Hermansky, Ellis, Doddington, Chen, Cretin, Bourlard, Athineos, "Pushing the envelope - aside [speech recognition]," IEEE Signal Processing Magazine, vol. 22, no. 5, 2005. — 2005

← **DARPA EARS Program 2001-2004: Novel Approach I**

**Bottle-neck Features extracted from Neural-Nets**

- Grezl, Karafiat, Kontar & Cernocky. "Probabilistic and bottle-neck features for LVCSR of meetings," ICASSP, 2007. — 2007

# Historical Development and Future Directions in Speech Recognition and Understanding

Janet M. Baker, Li Deng, Sanjeev Khudanpur, Chin-Hui Lee, James Glass, and Nelson Morgan

*This report is one of five reports that were based on the MINDS workshops, led by Donna Harman (NIST) and sponsored by Heather McCallum-Bayliss of the Disruptive Technology Office of the Office of the Director of National Intelligence's Office of Science and Technology (ODNI/ADDNI/S&T/DTO). To find the rest of the reports, and an executive overview, please see http://www.itl.nist.gov/iaui/894.02/minds.html.*

..., ..., ...

## 3. Models, Algorithms, and Search

**Machine Learning:** This is an exciting time in the machine learning community. Many new machine-learning algorithms are being explored and are achieving impressive results on a wide variety of tasks. Recent examples include graphical models, conditional random fields, (partially observable) Markov decision processes, reinforcement-based learning and discriminative methods such as large-margin or log-linear (max entropy) models. Recent developments in effective training of these models make them worthy of further exploration. The speech community would do well to explore common ground with the machine learning community in these areas.

# Deep Generative Models in ASR
## (prior to the rising of deep learning)

### Segment & Nonstationary-State Models

- Digalakis, Rohlicek, Ostendorf. "ML estimation of a stochastic linear system with the EM alg & application to speech recognition," IEEE T-SAP, 1993 — 1993
- Deng, Aksmanovic, Sun, Wu, Speech recognition using HMM with polynomial regression functions as nonstationary states," IEEE T-SAP, 1994. — 1994

### Hidden Dynamic Models (HDM)

- Deng, Ramsay, Sun. "Production models as a structural basis for automatic speech recognition," Speech Communication, vol. 33, pp. 93–111, 1997. — 1997
- Bridle et al. "An investigation of segmental hidden dynamic models of speech coarticulation for speech recognition," Final Report Workshop on Language Engineering, Johns Hopkins U, 1998. — 1998
- Picone et al. "Initial evaluation of hidden dynamic models on conversational speech," ICASSP, 1999. — 1999
- Deng and Ma. "Spontaneous speech recognition using a statistical co-articulatory model for the vocal tract resonance dynamics," JASA, 2000. — 2000

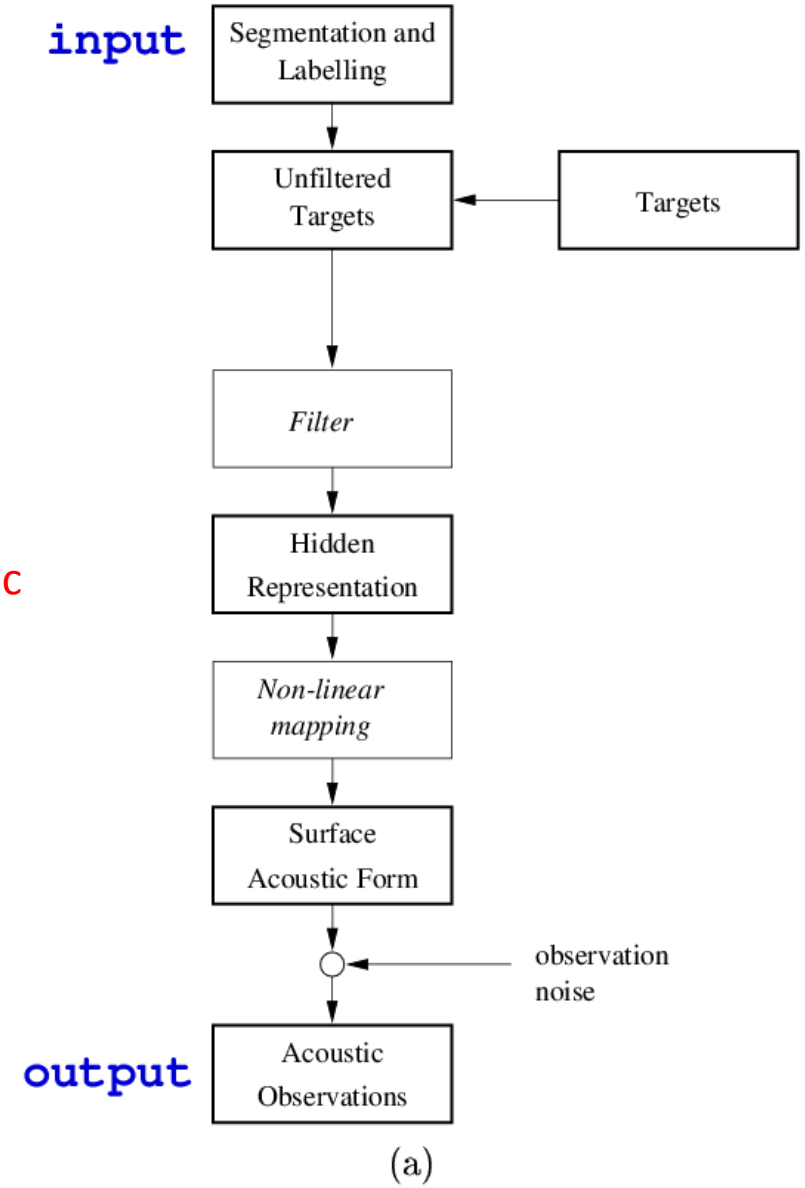
### Structured Hidden Trajectory Models

- Zhou, et al. "Coarticulation modeling by embedding a target-directed hidden trajectory model into HMM," ICASSP, 2003. ← **DARPA EARS Program 2001-2004: Novel Approach II** — 2003
- Deng, Yu, Acero. "Structured speech modeling," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 5, 2006. — 2006

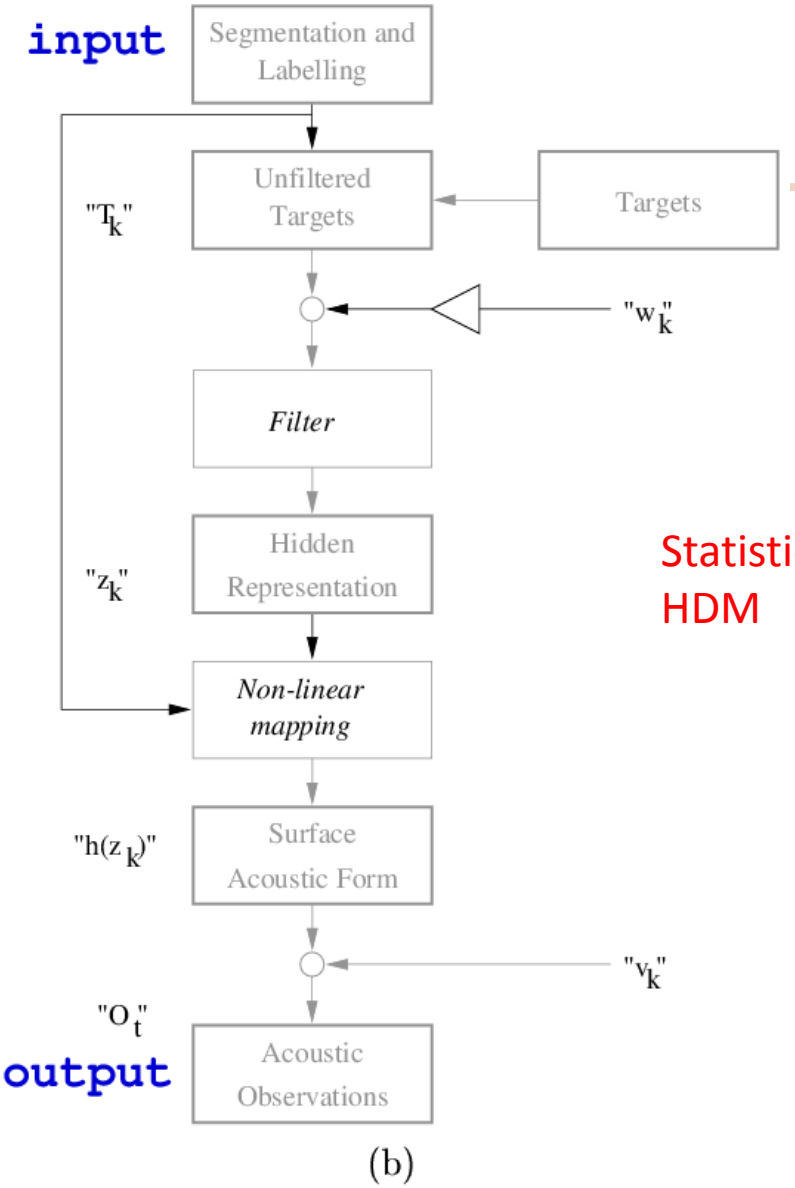
### Switching Nonlinear State-Space Models

- Deng. "Switching Dynamic System Models for Speech Articulation and Acoustics," in Mathematical Foundations of Speech and Language Processing, vol. 138, pp. 115 - 134, Springer, 2003.
- Lee et al. "A Multimodal Variational Approach to Learning and Inference in Switching State Space Models," ICASSP, 2004.

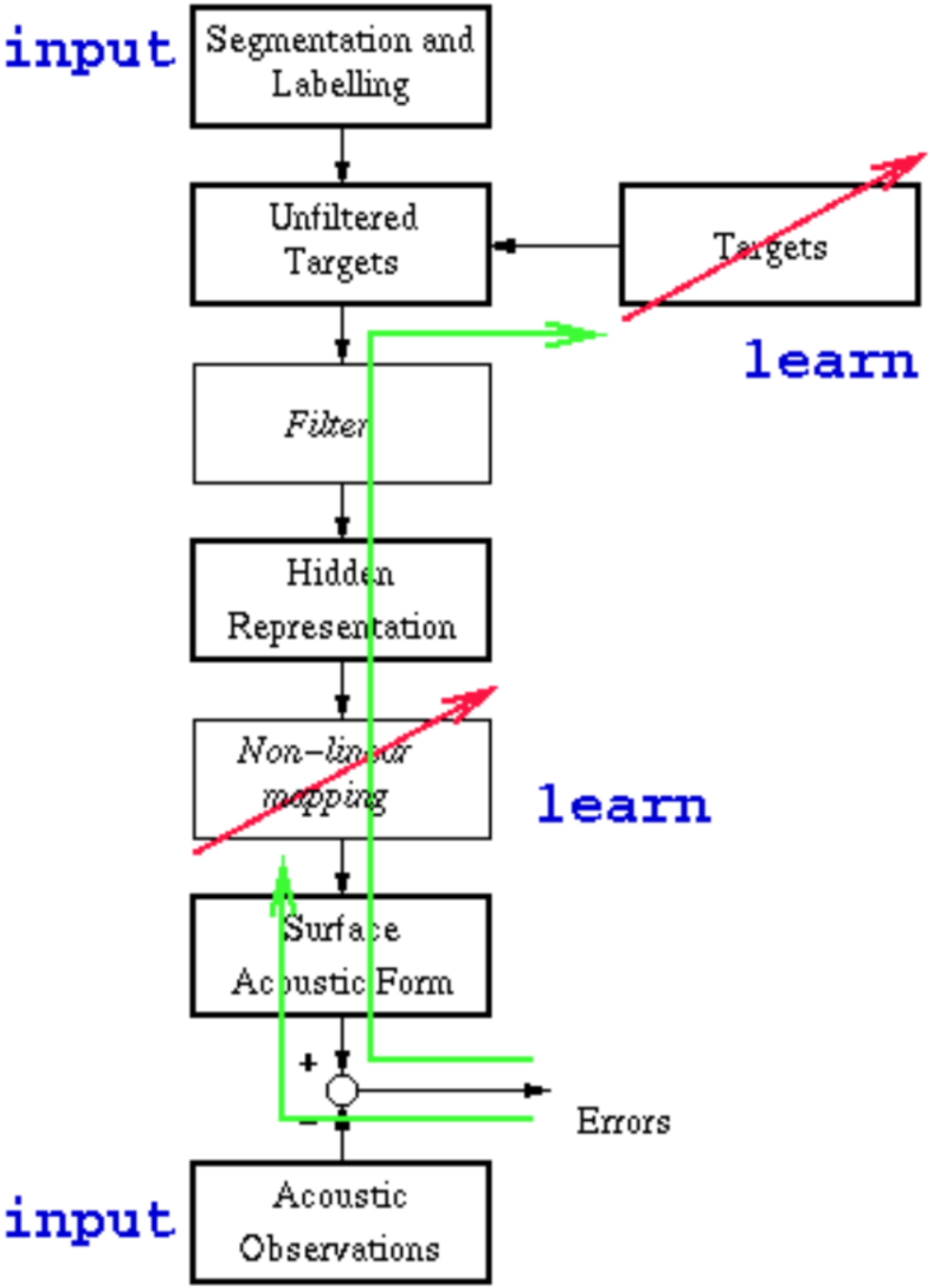
Deterministic HDM




Statistical HDM




Bridle, Deng, Picone, Richards, Ma, Kamm, Schuster, Pike, Reagan. Final Report for Workshop on Language Engineering, Johns Hopkins University, 1998.



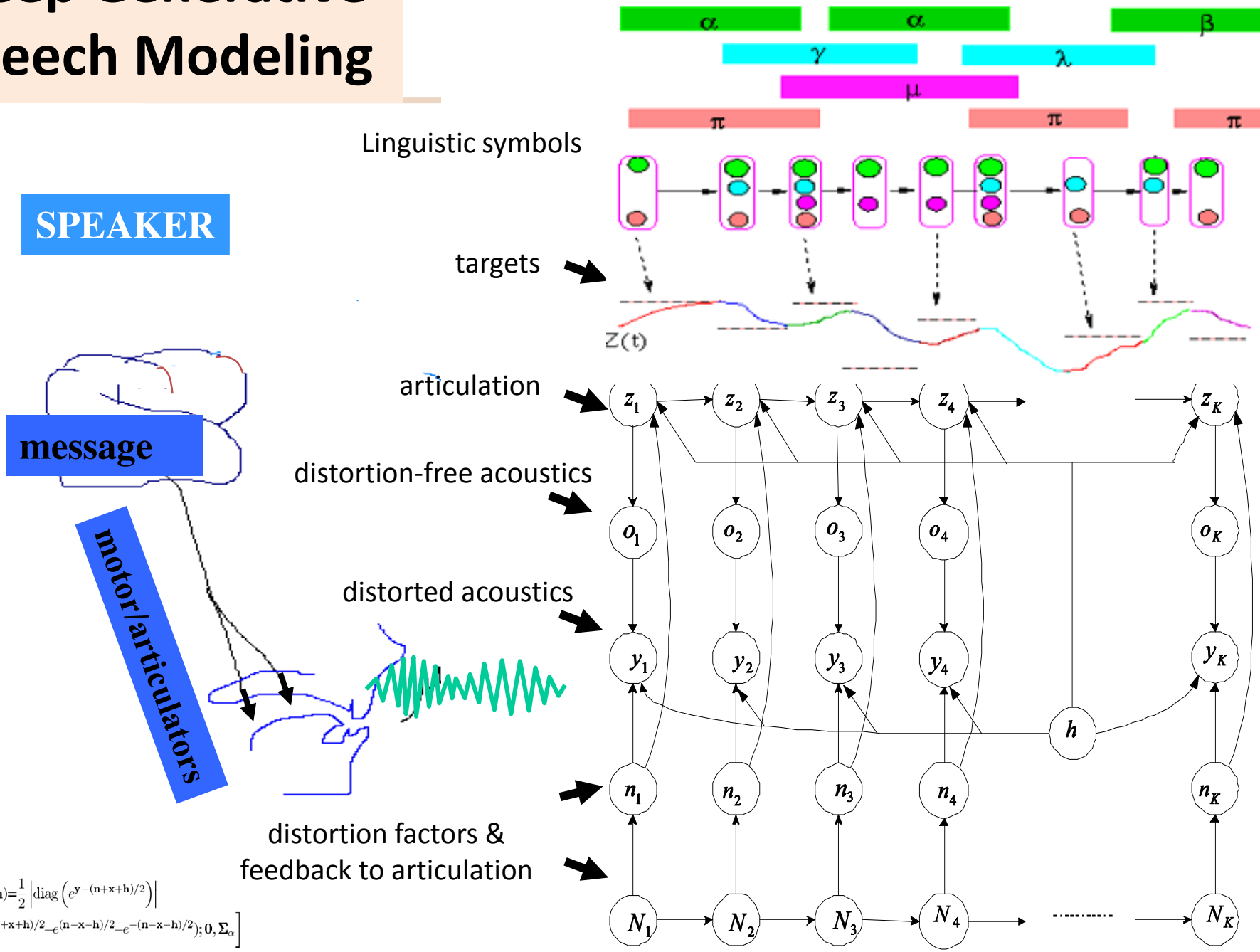

- Learning: Gradient descent
- "Back"-propagation in forward direction (model is generative)
- MSE for BP: acoustic obs, not labels
- Everything is interpretable
- Evaluation on SWBD: disappointment
- We understand why now, not in 1998
- Part II to discuss deep RNN vs. HDM



http://old-site.clsp.jhu.edu/ws98/projects/dynamic/presentations/hywel/presentation/training.html

# Deep Generative Speech Modeling

Ostendorf. “Moving beyond the ‘beads-on-a-string’ model of speech,” ASRU, 1999
Deng, Wu, “Hierarchical partitioning of the articulatory state space …,” ICSLP, 1996




$$p_y(\mathbf{y}|\mathbf{x},\mathbf{n},\mathbf{h}) = \frac{1}{2}\left|\mathrm{diag}\left(e^{\mathbf{y}-(\mathbf{n}+\mathbf{x}+\mathbf{h})/2}\right)\right| \mathcal{N}\left[\frac{1}{2}(e^{\mathbf{y}-(\mathbf{n}+\mathbf{x}+\mathbf{h})/2} - e^{(\mathbf{n}-\mathbf{x}-\mathbf{h})/2} - e^{-(\mathbf{n}-\mathbf{x}-\mathbf{h})/2}); \mathbf{0}, \mathbf{\Sigma}_\alpha\right]$$

ICASSP-2004

# A MULTIMODAL VARIATIONAL APPROACH TO LEARNING AND INFERENCE IN SWITCHING STATE SPACE MODELS

*Leo J. Lee*[1,2], *Hagai Attias*[2], *Li Deng*[2] *and Paul Fieguth*[3]

University of Waterloo
[1]Electrical & Computer Engineering
[3]Systems Design Engineering
Waterloo, ON, N2L 3G1
Canada

[2]Microsoft Corporation
Microsoft Research
One Microsoft Way
Redmond, WA 98052-6339
USA

Auxiliary function:

$$\mathcal{F}[q] = \sum_{s_{1:N}} \int d\mathbf{x}_{1:N}\; q(s_{1:N}, \mathbf{x}_{1:N}) \cdot [\log p(\mathbf{y}_{1:N}, \mathbf{x}_{1:N}, s_{1:N}) - \log q(s_{1:N}, \mathbf{x}_{1:N})]$$

In the variational approach we approximate the exact posterior $p(s_{1:N}, \mathbf{x}_{1:N} \mid \mathbf{y}_{1:N})$ by a distribution with a tractable structure, denoted by $q$. Here we choose the following partially factorized structure shown graphically in Fig. 1:

$$p(s_{0:N}, \mathbf{x}_{0:N} \mid \mathbf{y}_{1:N}) \approx q(s_{0:N}, \mathbf{x}_{0:N} \mid \mathbf{y}_{1:N}) = \prod_{n=1}^{N} q(\mathbf{x}_n \mid s_n) q(s_n \mid s_{n-1}) \cdot q(x_0 \mid s_0) q(s_0) \; . \tag{5}$$

**E-step: sufficient statistics.** As usual, the variational equations above are coupled, with the equations for $\rho_{s,n}$, $\Gamma_{s,n}$ depend on $\eta_{s's,n}$, $\gamma_{s,n}$ and vice versa. These equations are solved iteratively starting from a random or more suitable initialization if available. The solution is the set of sufficient statistics

$$\varphi = \{\rho_{s,n}, \Gamma_{s,n}, \eta_{ss',n}, \gamma_{s,n}\} \tag{16}$$

which are moments of the variational posterior.

**M-step: parameter estimation.** Given the sufficient statistics $\varphi$, the derivation of the M-step is achieved by taking derivatives of $\mathcal{F}$ w.r.t. the model parameters (details omitted).

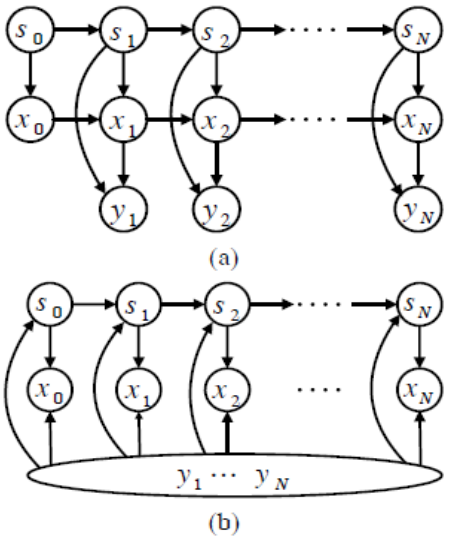
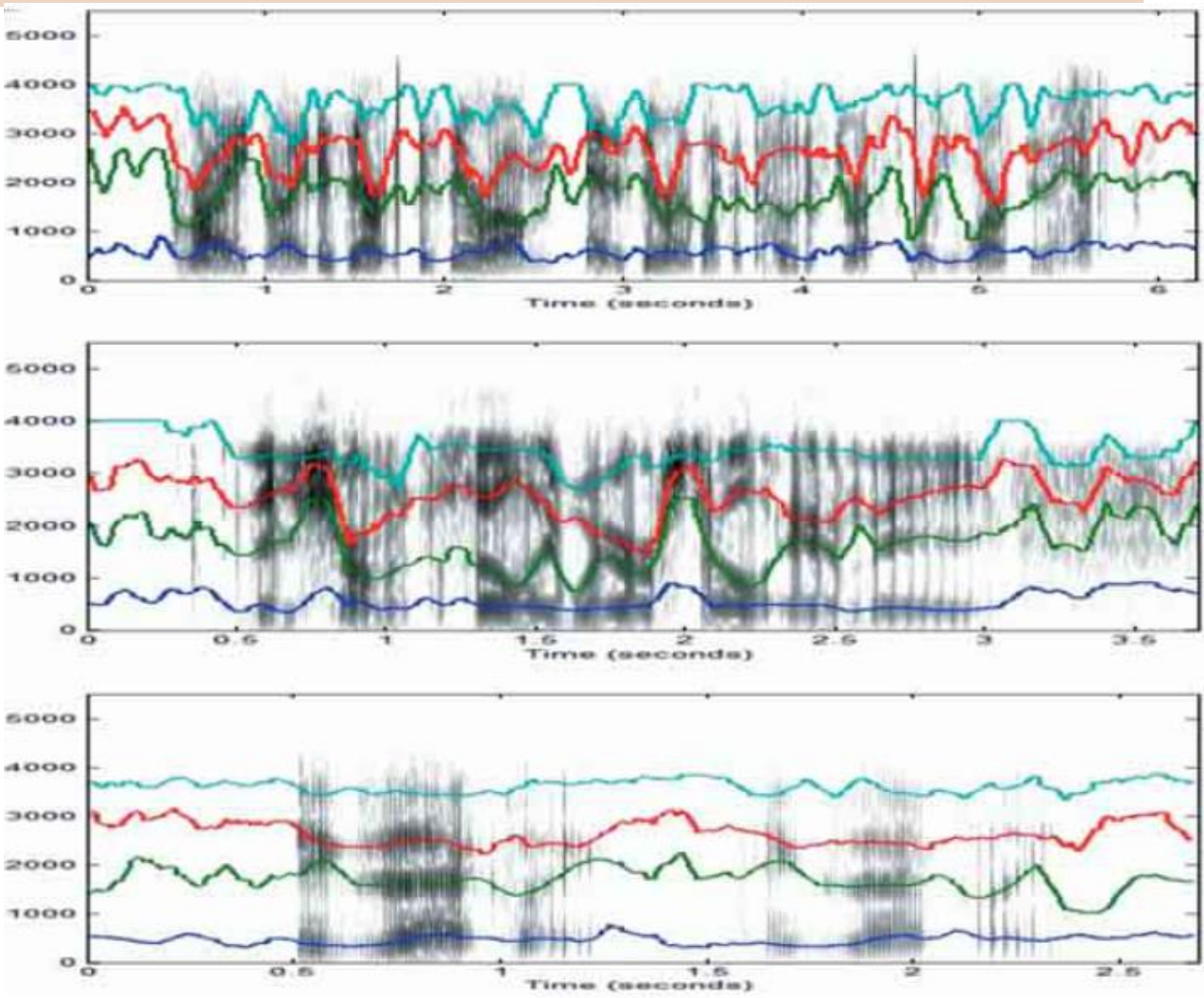



**Fig. 1**. The model (a) and the variational posterior (b) represented as Bayesian networks.

# Surprisingly Good Inference Results for Continuous Hidden States

- By-product: accurately tracking dynamics of resonances (formants) in vocal tract (TIMIT & SWBD).
- Best formant tracker (speech analysis); used as basis to form a formant database as "ground truth"
- We thought we solved the ASR problem, except
- "Intractable" for decoding




Deng & Huang, Challenges in Adopting Speech Recognition, *Communications of the ACM*, vol. 47, pp. 69-75, 2004.
Deng, Cui, Pruvenok, Huang, Momen, Chen, Alwan, A Database of Vocal Tract Resonance Trajectories for Research in Speech , ICASSP, 2006.

# Structured Speech Modeling

Li Deng, *Fellow, IEEE*, Dong Yu, *Member, IEEE*, and Alex Acero, *Fellow, IEEE*

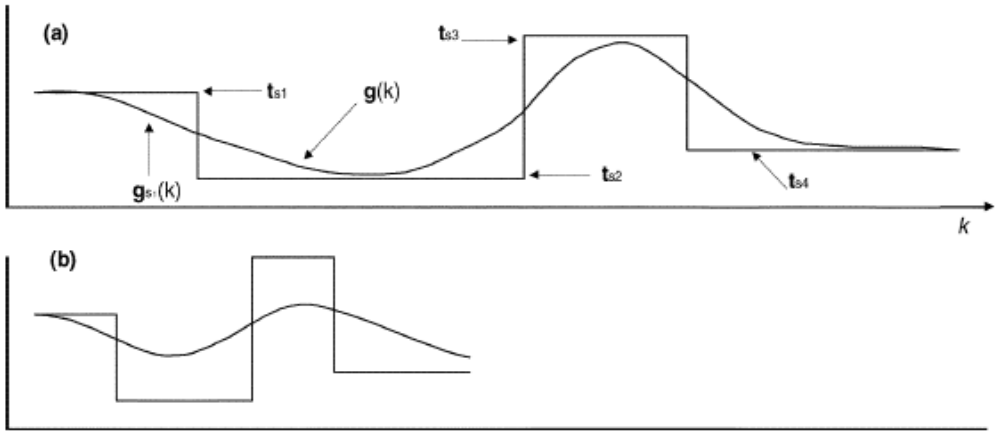



Fig. 1. Illustrations of the various VTR quantities in model Stage-I in utterance with four phone segments. (a) and (b) are for the same four VT targets and their filtered results, but the durations of the four segments a shorter in (b) than in (a).

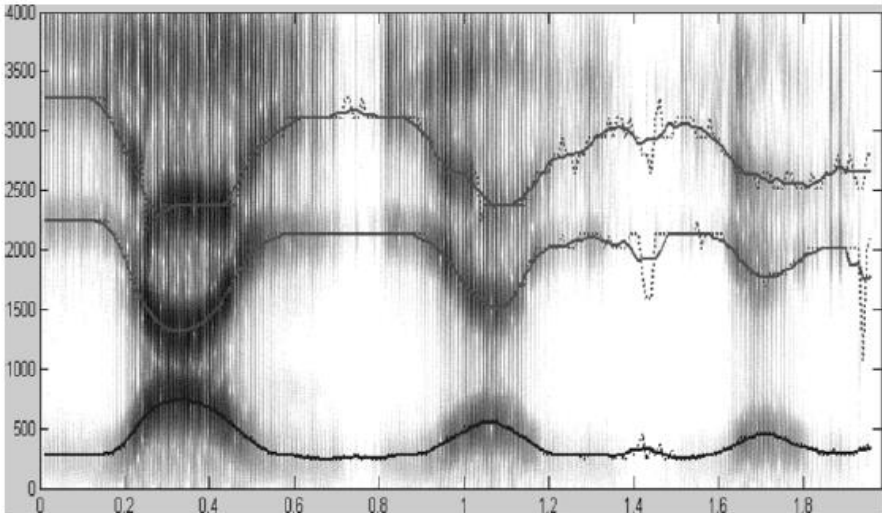



TABLE II
COMPARISONS OF HMM AND HTM PERFORMANCES (PERCENT CORRE
WITHIN EACH OF FOUR BROAD PHONE CLASSES

| | Sonorants | Stops | Fricatives | Closures |
|---|---|---|---|---|
| Occurrences | 3814 | 889 | 1252 | 1578 |
| HMM | 64.05 | 72.10 | 75.64 | 88.72 |
| HTM | 72.42 | 76.27 | 75.74 | 90.94 |

- Elegant model formulation & knowledge incorporation
- Strong empirical results: **96%** TIMIT accuracy with Nbest=1001; **75.2%** lattice decoding w. monophones; fast approx. training
- Still very expensive for decoding; could not ship (very frustrating!)

--corrected many errors:
especially “short” phones
--easy to interpret from the
“generative” story

# Another Deep Generative Model (developed outside speech)

- Sigmoid belief nets & wake/sleep alg. (1992)
- Deep belief nets (DBN, 2006);

→ Start of deep learning

- Totally non-obvious result:
  Stacking many RBMs (undirected)

→ not Deep Boltzmann Machine (**DBM**, undirected)

→ but a DBN (directed, generative model)

- Excellent in generating images & speech synthesis

- Similar type of deep generative models to HDM
- But simpler: no temporal dynamics
- With very different parameterization
- Most intriguing of DBN: inference is easy
  (i.e. no need for approximate variational Bayes)
  ← "Restriction" of connections in RBM

- Pros/cons analysis→ Hinton coming to MSR 2009

*LETTER* Communicated by Yann Le Cun

**A Fast Learning Algorithm for Deep Belief Nets**

**Geoffrey E. Hinton**
*hinton@cs.toronto.edu*
**Simon Osindero**
*osindero@cs.toronto.edu*
*Department of Computer Science, University of Toronto, Toronto, Canada M5S 3G4*

**Yee-Whye Teh**
*tehyw@comp.nus.edu.sg*
*Department of Computer Science, National University of Singapore, Singapore 117543*

We show how to use "complementary priors" to eliminate the explaining-away effects that make inference difficult in densely connected belief nets that have many hidden layers. Using complementary priors, we derive a fast, greedy algorithm that can learn deep, directed belief networks one layer at a time, provided the top two layers form an undirected associative memory. The fast, greedy algorithm is used to initialize a slower learning procedure that fine-tunes the weights using a contrastive version of the wake-sleep algorithm. After fine-tuning, a network with three hidden layers forms a very good generative model of the joint distribution of handwritten digit images and their labels. This generative model gives better digit classification than the best discriminative learning algorithms. The low-dimensional manifolds on which the digits lie are modeled by long ravines in the free-energy landscape of the top-level associative memory, and it is easy to explore these ravines by using the directed connections to display what the associative memory has in mind.

# then Geoff Hinton came to MSR (2009)

- **Kluge 1:** keep the assumption of "frame" independence (i.e. ignore real "dynamics" to facilitate inference/decoding) but use bigger time windows to approximate the effects.
- **Kluge 2:** reverse the direction: instead of "deep generating" speech top-down, do "deep inference" bottom-up (using neural nets → no problem of "explaining away")
- **Kluge 3:** don't know how to train this deep neural net? Try DBN to initialize it.
- **Well-timed** academic-industrial collaboration:
  - ASR industry searching for new solutions when "principled" deep generative approaches could not deliver
  - Academia developed deep learning tools (**DBN**/DNN with hybrid generative/discriminative, 2006) looking for applications
  - Advent of GPU computing (Nvidia CUDA library released 2007/08)
  - Big training data in ASR were available




Li Deng, Dong Yu, Geoffrey Hinton

Microsoft Research; Microsoft Research; University of Toronto

## Deep Learning for Speech Recognition and Related Applications

7:30am - 6:30pm Saturday, December 12, 2009

**Location:** Hilton: Cheakamus

**Abstract:** Over the past 25 years or so, speech recognition technology has been dominated by a "shallow" architecture --- hidden Markov models (HMMs). Significant technological success has been achieved using complex and carefully engineered variants

Invitee 1: give me one week to decide …,…

Not worth my time to fly to Vancouver for this…

there has been virtually no effective communication between machine learning researchers and speech recognition researchers who are both advocating the use of deep architecture and learning. One goal of the proposed workshop is to bring together these two groups of researchers to review the progress in both fields and to identify promising and synergistic research directions for potential future cross-fertilization and collaboration.

**Mohamed, Dahl, Hinton, Deep belief networks for phone recognition, *NIPS 2009 Workshop on Deep Learning,* 2009**

Yu, Deng, Wang, Learning in the Deep-Structured Conditional Random Fields, *NIPS 2009 Workshop on Deep Learning,* 2009

…, …, …


NIPS Home
Overview
Conference Videos
Workshop Videos
Program Highlights
Tutorials
Conference Sessions
Workshops
Publication Models
Demonstrations
Mini Symposia
Accepted Papers
Dates
Committees
Sponsors
Awards
Board

Li Deng, Dong Yu, Geoffrey Hinton

Microsoft Research; Microsoft Research; University of Toronto

**Deep Learning for Speech Recognition and Related Applications**

7:30am - 6:30pm Saturday, December 12, 2009

**Location:** Hilton: Cheakamus

**Abstract:** Over the past 25 years or so, speech recognition technology has been dominated by a "shallow" architecture --- hidden Markov models (HMMs). Significant

Invitee 2: A crazy idea…
Waveform for ASR is not like pixels for image recognition. It is more like using photons!!!

theoretical guidance to facilitate the development of these deep architectures. Further, there has been virtually no effective communication between machine learning researchers and speech recognition researchers who are both advocating the use of deep architecture and learning. One goal of the proposed workshop is to bring together these two groups of researchers to review the progress in both fields and to identify promising and synergistic research directions for potential future cross-fertilization and collaboration.

**Mohamed, Dahl, Hinton, Deep belief networks for phone recognition, *NIPS 2009 Workshop on Deep Learning,* 2009**
Yu, Deng, Wang, Learning in the Deep-Structured Conditional Random Fields, *NIPS 2009 Workshop on Deep Learning,* 2009
…, …, …

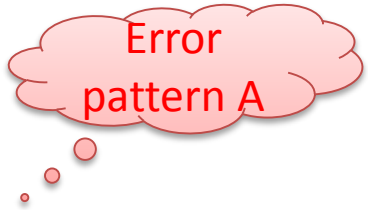
# Outline

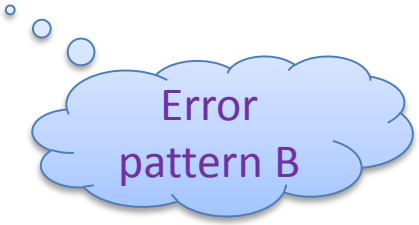
- **Part I:** A (brief) early history of 'deep' speech recognition
- **Part II:** Deep learning achievements: **speech** & vision
- **Part III:** Deep learning challenges: **Natural language**, **multimodality**, reasoning, mind, & deep intelligence in the big-data world

# A key discovery at MSR, 2009-2010

Error rates for the TIMIT phone recognition task:

| | |
|---|---|
| **Hidden Dynamic/Trajectory Model** | **24.8%** |
| Deep Neural Network | 23.4% |




Error-pattern-A

IS VERY DIFFERENT FROM

Error-pattern-B

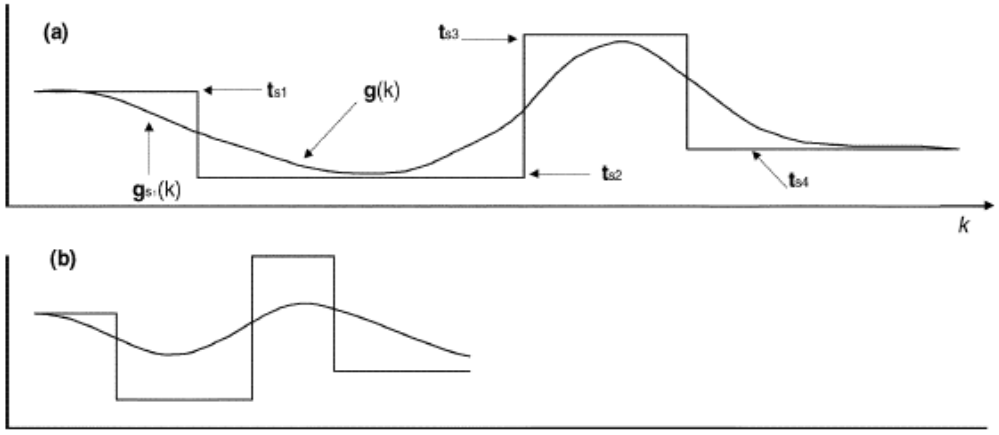



Fig. 1. Illustrations of the various VTR quantities in model Stage-I in an utterance with four phone segments. (a) and (b) are for the same four VTR targets and their filtered results, but the durations of the four segments are shorter in (b) than in (a).

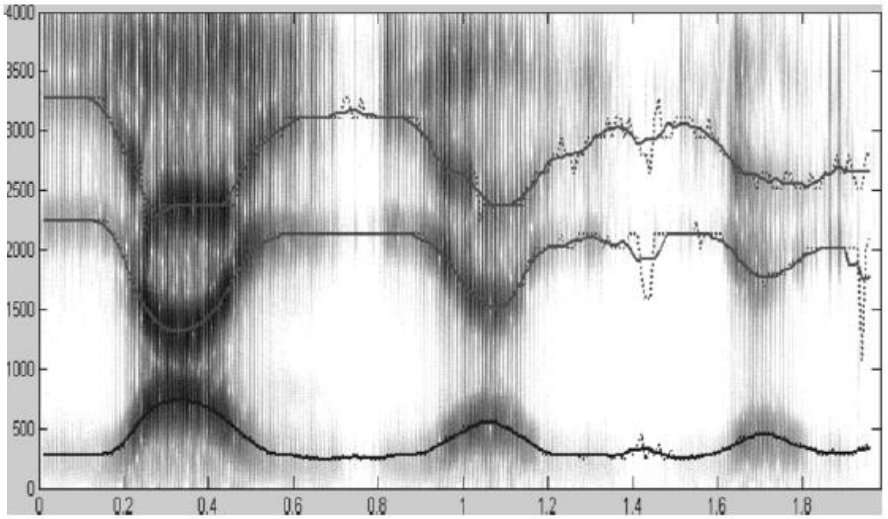



Fig. 2. Spectrogram of three renditions of /iy aa iy/ by one author, with an increasingly higher speaking rate and increasingly lower speaking efforts. The horizontal label is time, and the vertical one is frequency.

TABLE II
COMPARISONS OF HMM AND HTM PERFORMANCES (PERCENT CORRECT) WITHIN EACH OF FOUR BROAD PHONE CLASSES

| | Sonorants | Stops | Fricatives | Closures |
|---|---|---|---|---|
| Occurrences | 3814 | 889 | 1252 | 1578 |
| HMM | 64.05 | 72.10 | 75.64 | 88.72 |
| HTM | 72.42 | 76.27 | 75.74 | 90.94 |

-- DNN made many new errors
on short, undershoot vowels
-- 11 frames contain too much “noise”

# Another key discovery at MSR, 2009-2010

- Spectrogram (fbank) features better than cepstra features (MFCCs)

  (on speech analysis & feature coding using deep autoencoders)
- MFCCs dominated speech analysis & recognition: 1982-2012
- Conforming to the basic DL theme: back to raw features (and learn transformations automatically layer by layer)

# Binary Coding of Speech Spectrograms Using a Deep Auto-encoder

*L. Deng[1], M. Seltzer[1], D. Yu[1], A. Acero[1], A. Mohamed[2], and G. Hinton[2]*

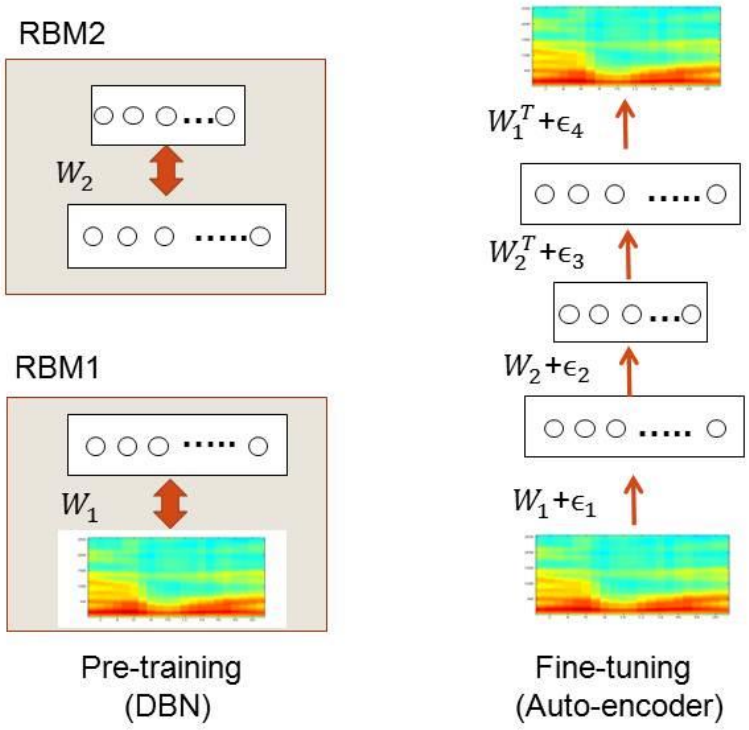
[1] Microsoft Research, One Microsoft Way, Redmond, WA 98052, US
[2] University of Toronto, Toronto, Ontario, Canada

{deng|mseltzer|dongyu|alexac}@microsoft.com; {asamir|hinton}@cs.toronto.edu

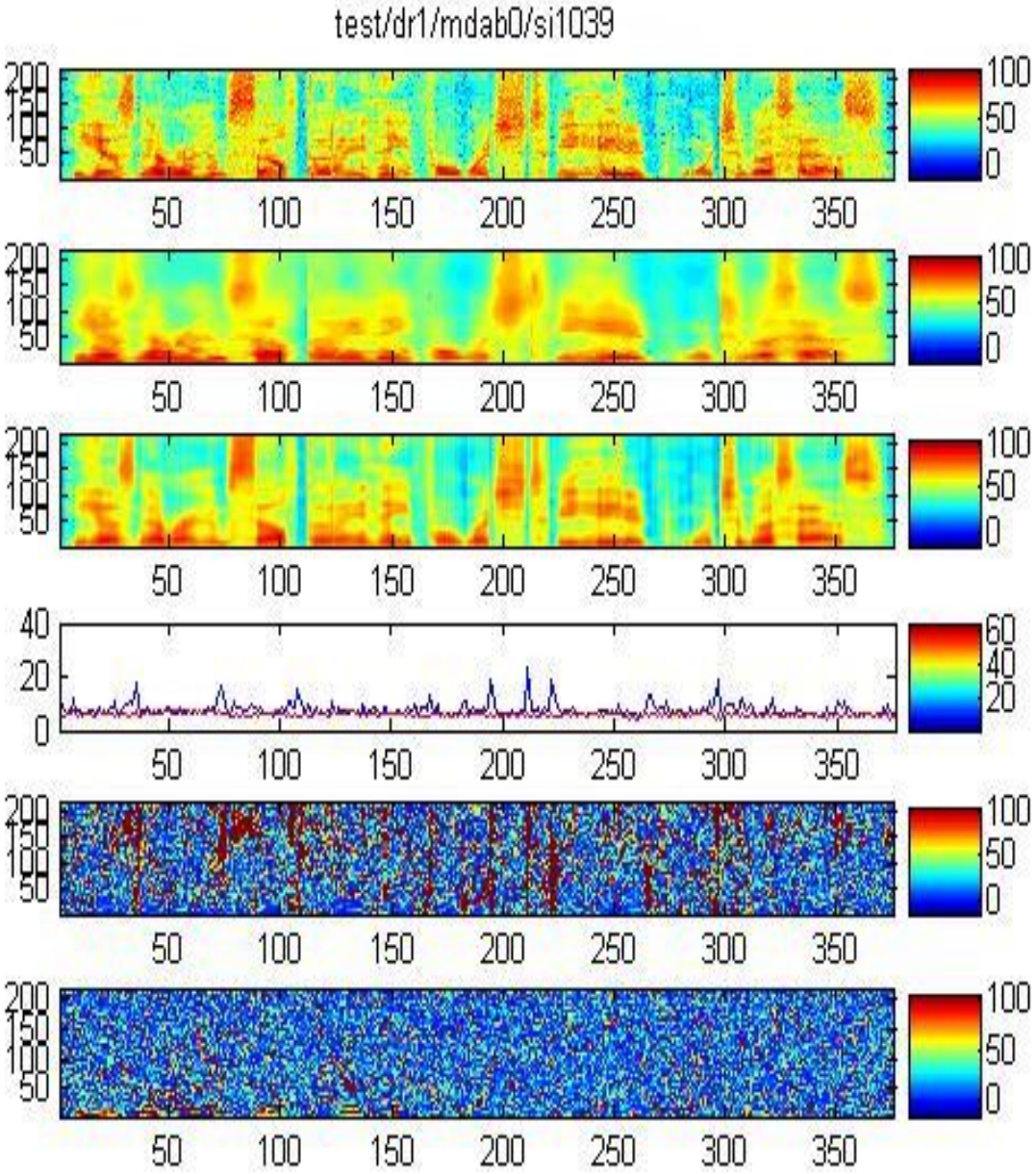
**Abstract**

This paper reports our recent exploration of the layer-by-layer

The work reported in this paper was inspired by the successful use of deep auto-encoders for dimensionality reduction [8][9] and the extension of this work to the




- First use of spectrogram features in speech analysis with deep learning
- Deep autoencoder better than "shallow" method of VQ
- Filterbank features better MFCCs




***Original spectrogram (log |FFT|)***

*Reconstructed spectrogram from a 312-bit VQ coder*

*Reconstructed spectrogram from a 312-bit **deep autoencoder***

*Coding errors as a function of time for VQ coder (blue) and autoencoder (red)*

*VQ coder's error (over time & freq)*

*Deep autoencoder's error*

- No such nice results for MFCCs

# Expanding DNN at Industry Scale

- Scale DNN's success to large speech tasks (2010-2011)
  - Grew output neurons from context-independent phone states (100-200) to context-dependent ones (1k-30k) → CD-DNN-HMM for Bing Voice Search and then to SWBD tasks
  - Motivated initially by saving huge MSFT investment in the speech decoder software infrastructure (several choices available: **senones**, symbolic articulatory "features", etc. )
  - CD-DNN-HMM gave much higher accuracy than CI-DNN-HMM
  - Earlier NNs made use of context only as appended inputs, not coded directly as outputs

- Engineering for large speech systems:
  - Combined expertise in DNN (esp. with GPU implementation) **and** speech recognition
  - Collaborations among MSRR/MSRA, academic researchers


- Yu, Deng, Dahl, Roles of Pre-Training and Fine-Tuning in Context-Dependent DBN-HMMs for Real-World Speech Recognition, in *NIPS Workshop on Deep Learning*, 2010.
- Dahl, Yu, Deng, Acero, Large Vocabulary Continuous Speech Recognition With Context-Dependent DBN-HMMS, in *Proc. ICASSP,*
- Dahl, Yu, Deng, Acero, Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition, in *IEEE Transactions on Audio, Speech, and Language Processing (2013 IEEE SPS Best Paper Award)* , vol. 20, no. 1, pp. 30-42, January 2012.
- Seide, Li, Yu, "Conversational Speech Transcription Using Context-Dependent Deep Neural Networks", Interspeech 2011, pp. 437-440.
- Hinton, Deng, Yu, Dahl, Mohamed, Jaitly, Senior, Vanhoucke, Nguyen, Sainath,Kingsbury, Deep Neural Networks for Acoustic Modeling in Speech Recognition, in *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82-97, November 2012

# What Enabled CD-DNN-HMM?

- Industry knowledge of
  - how to construct very large CD output units in DNN
  - how to make decoding of such huge networks highly efficient using HMM technology
  - how to cut corner in making practical systems
- GPUs are optimized for fast matrix multiplications, major computation in CD-DNN **training**
- Nvidia's CUDA library for GPU computing released in 2008

# DNN vs. Pre-DNN Prior-Art

- **Table:** TIMIT Phone recognition (3 hours of training)

| Features | Setup | Error Rates |
|---|---|---|
| Pre-DNN | Deep Generative Model | 24.8% |
| DNN | 5 layers x 2048 | 23.4% |

~10% relative improvement

- **Table:** Voice Search SER (24-48 hours of training)

| Features | Setup | Error Rates |
|---|---|---|
| Pre-DNN | GMM-HMM with MPE | 36.2% |
| DNN | 5 layers x 2048 | 30.1% |

~20% relative improvement

- **Table:** SwitchBoard WER (309 hours training)

| Features | Setup | Error Rates |
|---|---|---|
| Pre-DNN | GMM-HMM with BMMI | 23.6% |
| DNN | 7 layers x 2048 | 15.8% |

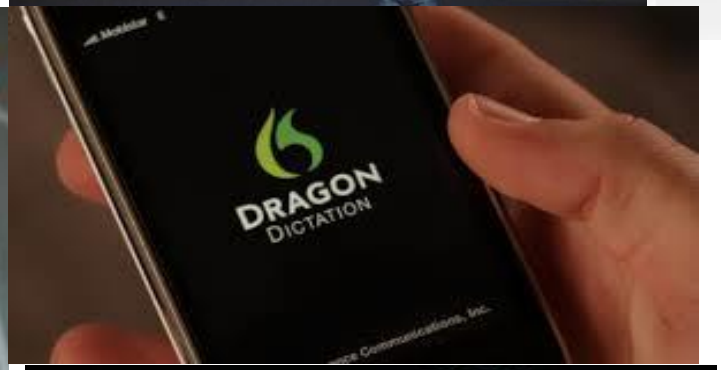
~30% relative Improvement (Seide et al, 2011)

For DNN, the more data, the better!

# Across-the-Board Deployment of DNN in ASR Industry

(+ university labs & DARPA program)






# Many, but NOT ALL, limitations of early DNNs have been overcome

- **Kluge 1:** keep the assumption of frame independence (ignore real "dynamics" to speed up decoding) but use bigger time windows

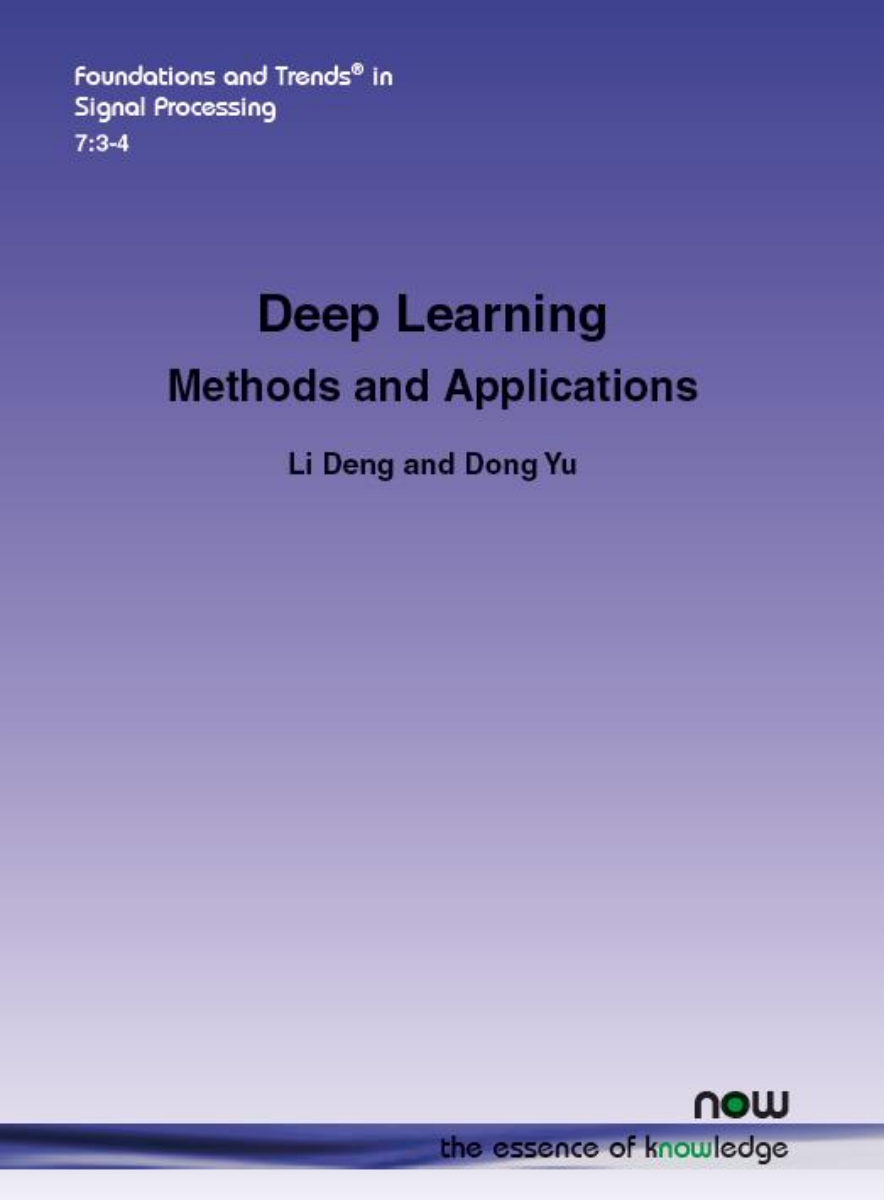
→ LSTM-RNN (single-frame input features)

- **Kluge 2:** reverse the direction: instead of "deep generating" speech top-down, do "deep inference" bottom-up (using neural nets)

→ NOT YET: integrating deep generative model and DNN

- **Kluge 3:** don't know how to train this deep neural net? Try DBN to initialize it.

→no need for DBN pre-training if you have big data; this is well understood now



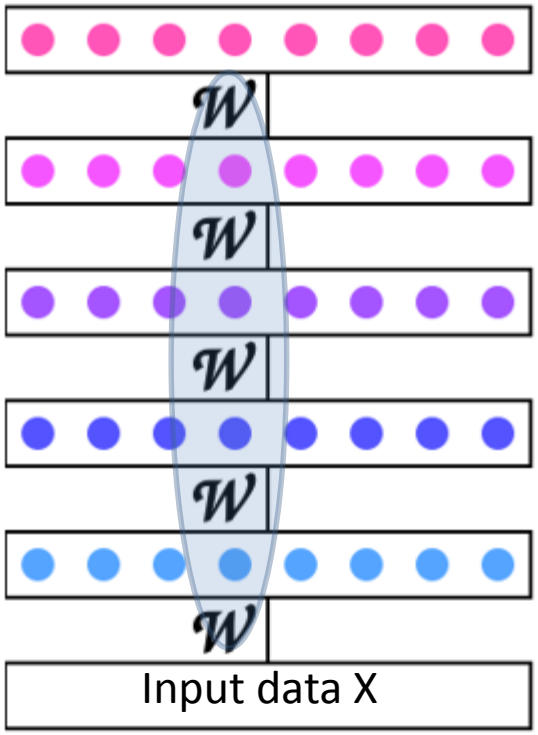

# Chapter 7

## Selected Applications in Speech and Audio Processing

### 7.1 Acoustic modeling for speech recognition

As discussed in Section 2, speech recognition is the very first successful application of deep learning methods at an industry scale. This success is a result of close academic-industrial collaboration, initiated at Microsoft Research, with the involved researchers identifying and acutely attending to the industrial need for large-scale deployment [68, 89, 109, 161, 323, 414]. It is also a result of carefully exploiting the strengths of the deep learning and the then-state-of-the-art speech recognition technology, including notably the highly efficient decoding techniques.

# Innovation: Better Optimization




- **Sequence discriminative training:**

- Mohamed, Yu, Deng: "Investigation of full-sequence training of deep belief networks for speech recognition," Interspeech, 2010.
- Kingsbury, Sainath, Soltau. "Scalable minimum Bayes risk training of DNN acoustic models using distributed hessian-free optimization," Interspeech, 2012.
- Su, Li, Yu, Seide. "Error back propagation for sequence training of CD deep networks for conversational speech transcription," ICASSP, 2013.
- Vesely, Ghoshal, Burget, Povey. "Sequence-discriminative training of deep neural networks, Interspeech, 2013.

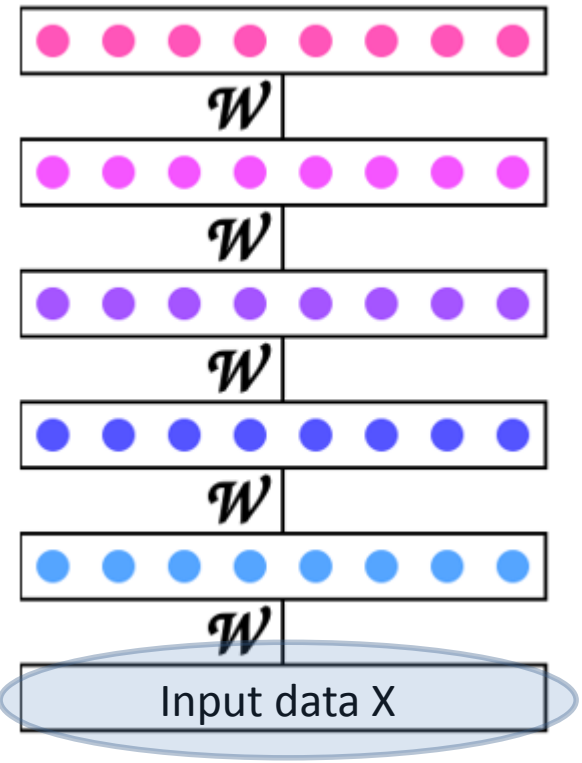
- **Distributed asynchronous SGD**

- Dean, Corrado,…Senior, Ng. "Large Scale Distributed Deep Networks," NIPS, 2012.
- Sak, Vinyals, Heigold, Senior, McDermott, Monga, Mao. "Sequence Discriminative Distributed Training of Long Short-Term Memory Recurrent Neural Networks," Interspeech,2014.

- **Primal-dual method**

- Chen, Deng. "A primal-dual method for training recurrent neural networks constrained by the echo-state property," ICLR, 2014.

# Innovation: Towards Raw Inputs




- **Bye-Bye MFCCs (Mel-scaling & cosine transform) !**

- Deng, Seltzer, Yu, Acero, Mohamed, Hinton. "Binary coding of speech spectrograms using a deep auto-encoder," Interspeech, **2010**.
- Mohamed, Hinton, Penn. "Understanding how deep belief networks perform acoustic modeling," ICASSP, 2012.
- Li, Yu, Huang, Gong, "Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM" SLT, 2012
- Deng, Li, Huang, Yao, Yu, Seide, Seltzer, Zweig, He, Williams, Gong, Acero. "Recent advances in deep learning for speech research at Microsoft," ICASSP, 2013.
- Sainath, Kingsbury, Mohamed, Ramabhadran. "Learning filter banks within a deep neural network framework," ASRU, 2013.

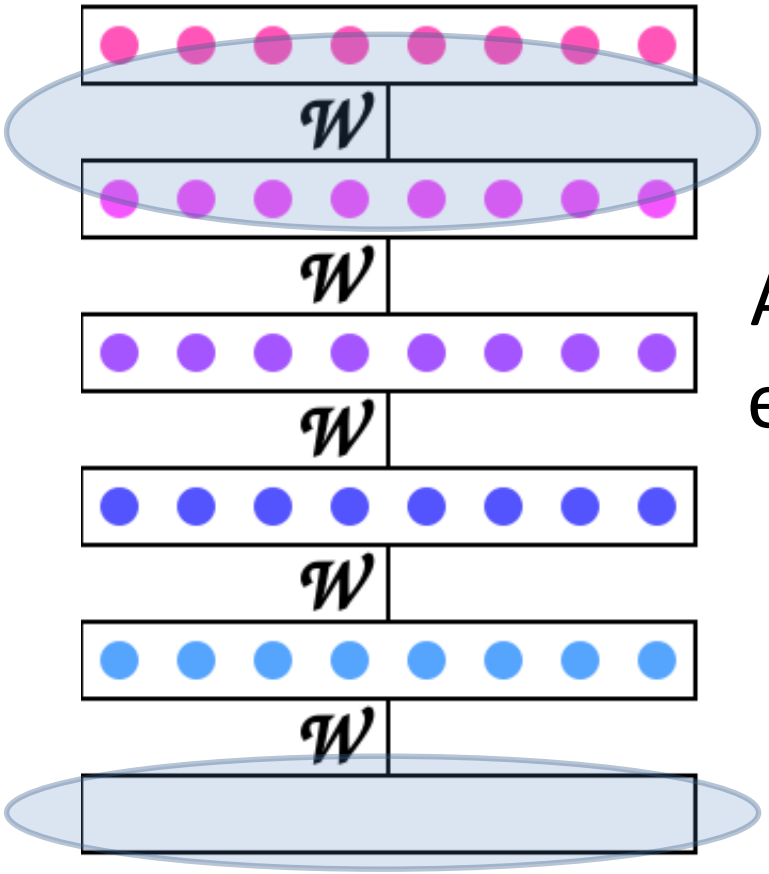
- **Bye-Bye Fourier transforms ?**

(- Sheikhzadeh, Deng, "Waveform-based speech recognition using hidden filter models," *IEEE T-SAP, 1994.)*
- Jaitly and Hinton. "Learning a better representation of speech sound waves using RBMs," ICASSP, 2011.
- Tuske, Golik, Schluter, Ney. "Acoustic modeling with deep neural networks using raw time signal for LVCSR," Interspeech, **2014**.

- **DNN as hierarchical nonlinear feature extractors:**

- Seide, Li, Chen, Yu. "Feature engineering in context-dependent deep neural networks for conversational speech transcription, ASRU, 2011.
- Yu, Seltzer, Li, Huang, Seide. "Feature learning in deep neural networks - Studies on speech recognition tasks," ICLR, 2013.
- Yan, Huo, Xu. "A scalable approach to using DNN-derived in GMM-HMM based acoustic modeling in LVCSR," Interspeech, 2013.
- Deng, Chen. "Sequence classification using high-level features extracted from deep neural networks," ICASSP, 2014.

# Innovation: Transfer/Multitask Learning & Adaptation



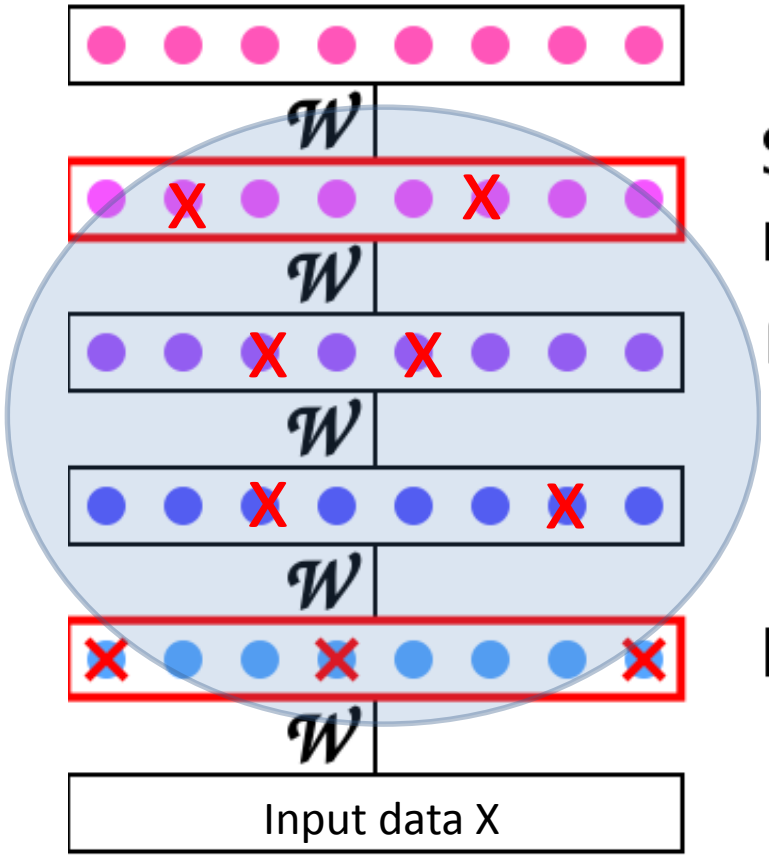

Multi-lingual acoustic modeling

Adaptation to speakers & environments

Mixed-bandwidth acoustic modeling

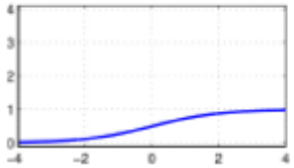
- Too many references to list & organize

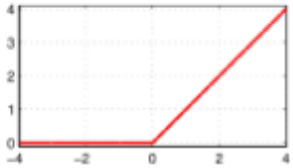
# Innovation: Better regularization & **nonlinearity**




Sparsity in hidden representations

logistic ⟶ ReLU , MaxOut,



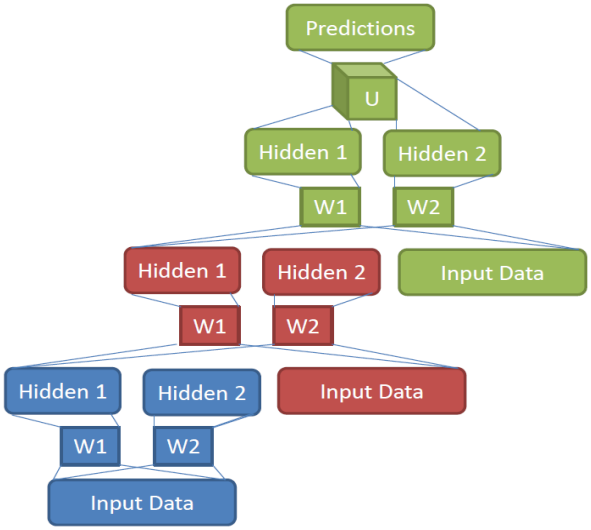


Dropout

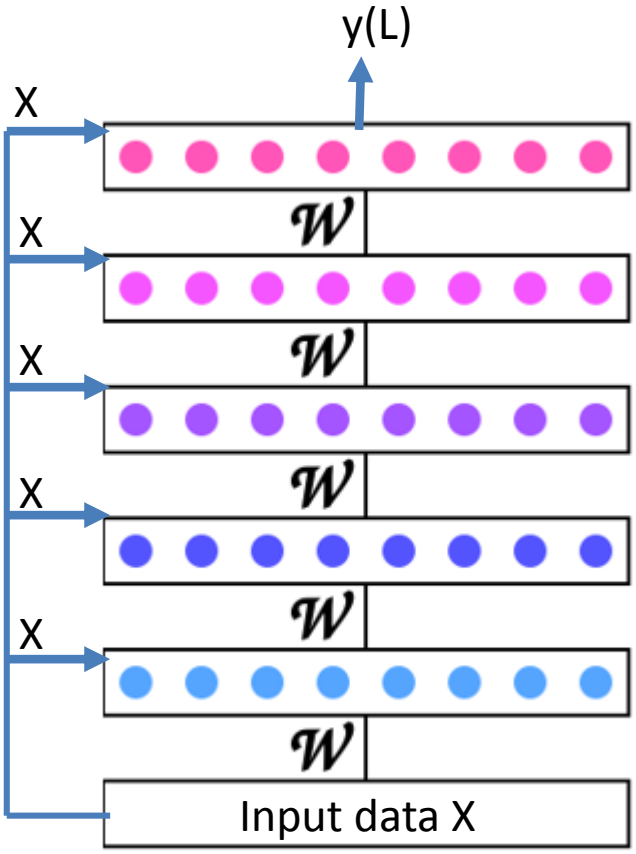
# Innovation: Better architectures

Another example: **Deep Stacking Network** & **Tensor DSN**

Deep Convex Network: A Scalable Architecture for Speech Pattern Classification, Interspeec
Scalable stacking and learning for building deep architectures, *ICASSP-2012*
Tensor Deep Stacking Networks, *IEEE T-PAMI,* 2013.
DEEP LEARNING --- Methods and Applications, 2014




Stacking w. hidden (or output) layer




$$y(\ell)=\mathcal{F}[y(\ell-1),\mathcal{W}(\ell-1),X]$$

$$y(\ell-1)=\mathcal{F}[y(\ell-2),\mathcal{W}(\ell-2),X]$$

$$\vdots$$

$$y(1)=\mathcal{F}[\mathcal{W}(1),X]$$

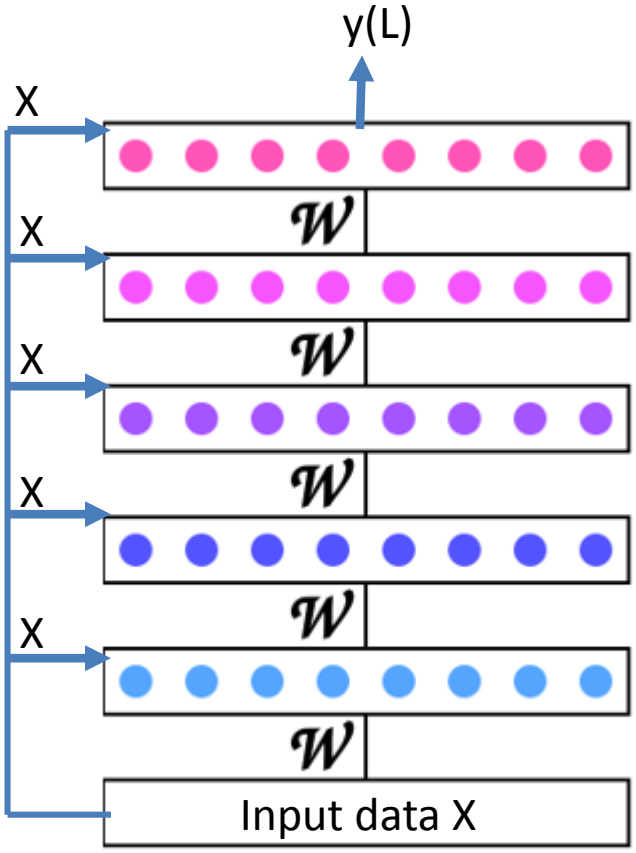
**Activation function $\mathcal{F}[\bullet]$**

- Pre-fixed
- Logistic or
- ReLu

# Innovation: Better architectures

Yet another example: **Deep unfolding network (inspired by generative modeling)**

-- Hershey, Le Roux, and Weninger, "Deep Unfolding: Model-Based Inspiration of Novel Deep Architectures," MERL TR2014-117 & ArXiv 2014.




$$\mathcal{F}[\bullet] = \mathbf{H}_t^k \circ \frac{(\mathbf{W}^k)^T \frac{\mathbf{M}_t}{\mathbf{W}\mathbf{H}_t^k}}{(\mathbf{W}^k)^T \mathbf{1} + \mu}.$$

$y(\ell)=\mathcal{F}[y(\ell\text{-}1), \mathcal{W}(\ell\text{-}1), X]$

$y(\ell\text{-}1)=\mathcal{F}[y(\ell\text{-}2), \mathcal{W}(\ell\text{-}2), X]$

$y(1)=\mathcal{F}[\mathcal{W}(1), X]$

- Activation function $\mathcal{F}[\bullet]$ derived from inference method in a generative model, not fixed in DSN
- The generative model embeds **knowledge/constraint** about how noisy speech is composed from clean speech & noise
  This (shallow) generative model, **non-negative matrix factorization**, unfolds in inference steps to form a DNN after untying weights
- Application: enhancement of speech & source sep. (demo)
- Example of how to integrate DNN with natural constraints in a shallow generative model
- How about deep generative model?

# Advances in Inference Algms for Deep Generative Models

Kingma 2013, Bengio 2013, Kingma & Welling 2014

**ICML-2014 Talk Monday June 23, 15:20**

**In Track F (Deep Learning II)**

**"Efficient Gradient Based Inference through Transformations between Bayes Nets and Neural Nets"**

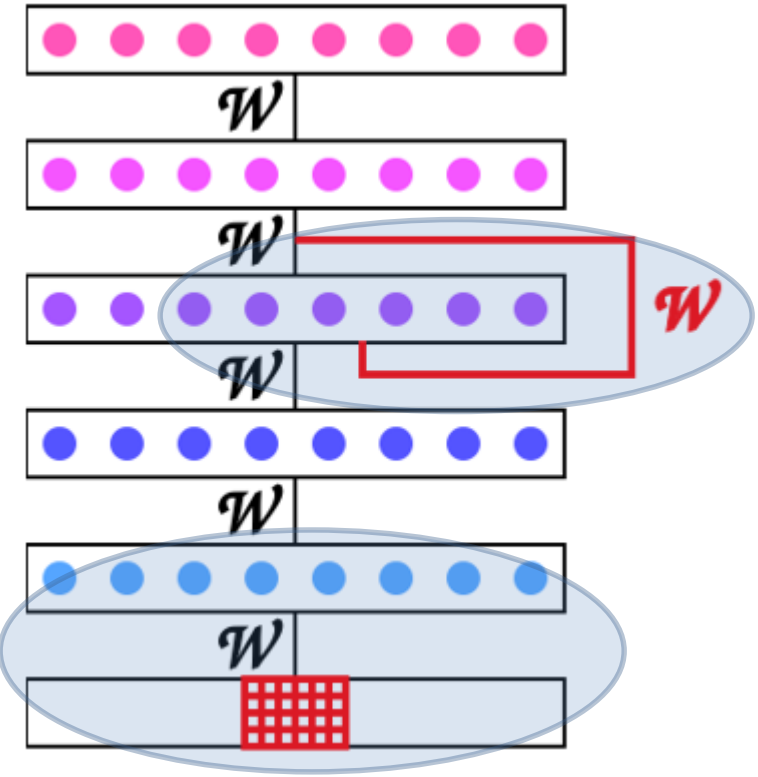
*Other solutions to solve the "large variance problem" in variational inference:*

-Variational Bayesian Inference with Stochastic Search [D.M. Blei, M.I. Jordan and J.W. Paisley, 2012]
-Fixed-Form Variational Posterior Approximation through Stochastic Linear Regression [T. Salimans and A. Knowles, 2013].
-Black Box Variational Inference. [R. Ranganath, S. Gerrish and D.M. Blei. 2013]
-Stochastic Variational Inference [M.D. Hoffman, D. Blei, C. Wang and J. Paisley, 2013]
-Estimating or **propagating gradients through stochastic neurons**. [Y. Bengio, 2013].
-Neural Variational Inference and Learning in Belief Networks. [A. Mnih and K. Gregor, 2014]



Slide provided by Max Welling (ICML-2014 tutorial) w. revision

# Innovation: Better architectures




- **Recurrent Nets (RNN) and Convolutional Nets (CNN) give state-of-the-art ASR results:**

- Sak, Senior, Beaufays. "LSTM Recurrent Neural Network architectures for large scale acoustic modeling," Interspeech,2014.
  → State-of-the-art results: **9.8% WER** for voice search

- Soltau, Saon, Sainath. "Joint Training of Convolutional and Non-Convolutional Neural Networks," ICASSP, 2014.
  → State-of-the-art results: **10.4% WER** for SWBD task (309hr training)

# Many, but NOT ALL, limitations of early DNNs have been overcome

- **Kluge 1:** keep the assumption of frame independence (ignore real "dynamics" to speed up decoding) but use bigger time windows

→ Just fixed by LSTM-RNN (single-frame input features)

- **Kluge 2:** reverse the direction: instead of "deep generating" speech top-down, do "deep inference" bottom-up (using neural nets)

→ NOT YET: integrating deep generative model and DNN

- **Kluge 3:** don't know how to train this deep neural net? Try DBN to initialize it.

→no need for DBN pre-training if you have big data;

this is well understood now

# Analyzing: RNN (no LSTM) vs. Generative HDM

$$\mathbf{h}_t = f(\mathbf{h}_{t-1}; \mathbf{W}_{hh}, \mathbf{W}_{xh,}, \mathbf{x}_t)$$
$$\mathbf{y}_t = g(\mathbf{h}_t; \mathbf{W}_{hy})$$

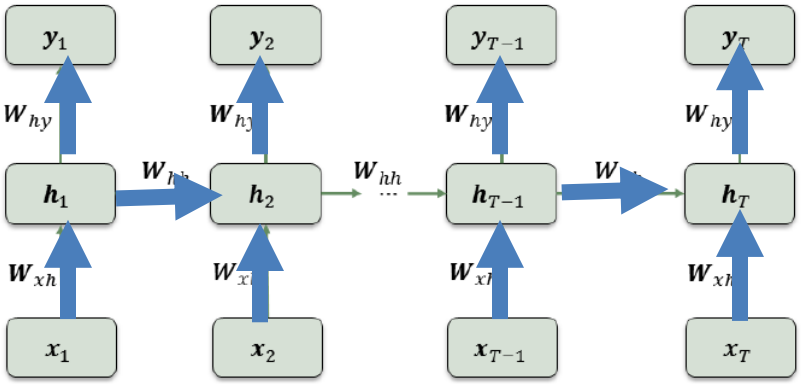
$$\mathbf{h}_t = q(\mathbf{h}_{t-1}; \mathbf{W}_{l_t}, \mathbf{t}_{l_t})$$
$$\mathbf{x}_t = r(\mathbf{h}_t, \Omega_{l_t})$$

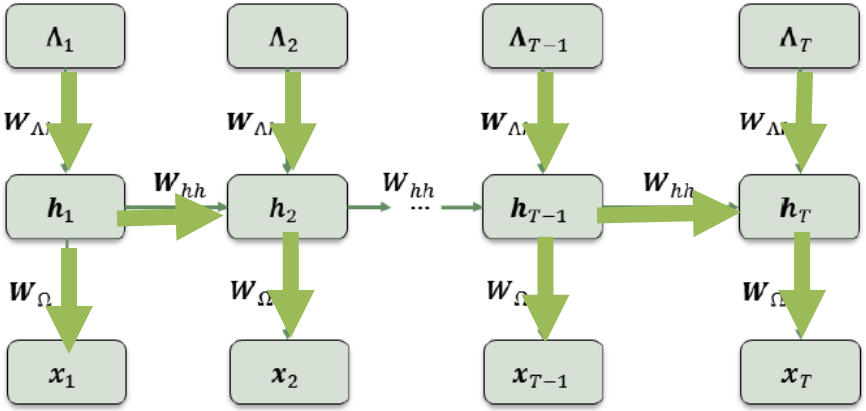
Parameterization:

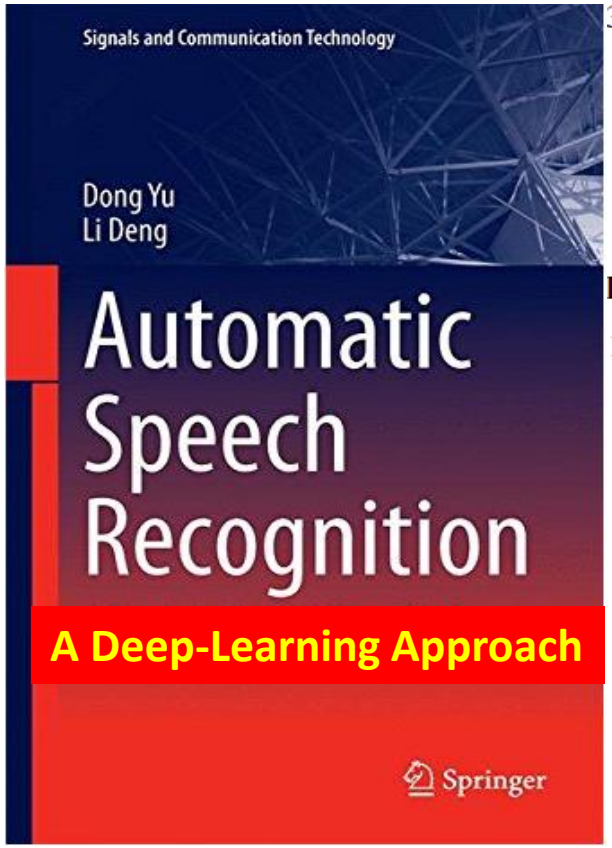
- $W_{hh}, W_{hy}, W_{xh}$: all unstructured regular matrices

Parameterization:

- $W_{hh}$ =M($\gamma_l$); sparse system matrix
- $W_\Omega$ =($\Omega_l$); Gaussian-mix params; MLP
- $\Lambda = \boldsymbol{t}_l$





# Outline

- Part I: A (brief) early history of "deep" speech recognition
- Part II: Deep learning achievements in speech and vision

- **Part III: Deep learning challenges: Language, multimodality,** mind, & deep intelligence
  - from DNN to deep semantic modeling (different problems & approaches)
  - **DSSM** developed at MSR for text/multimodal processing
  - functional modeling of the brain/mind for deep intelligence

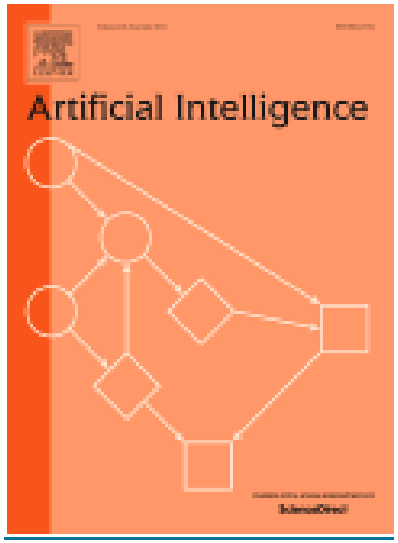
# A Key Concept

(IEEE/ACM Trans Audio Speech Language Proc., Special Issue, 2014)

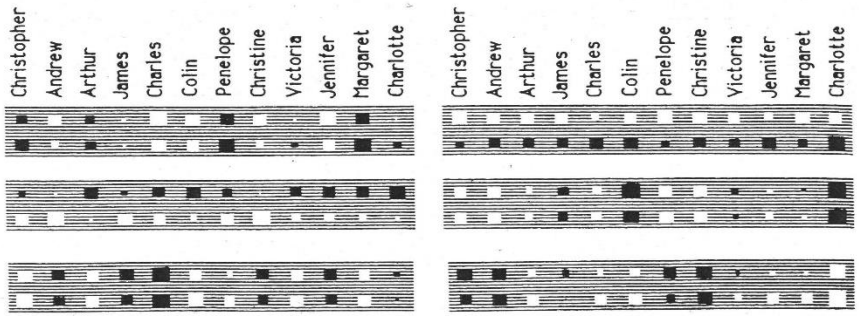
- **A linguistic or physical entity or a simple "relation"**

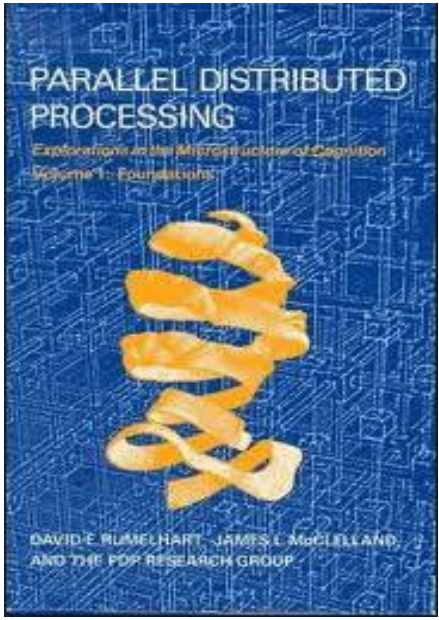
mapping via distributed representations by NN

**A low-dim continuous-space vector or embedding**




**Special Issue, vol. 46 (1990)**
**Connectionist Symbol Processing**
**(4 articles)**




**PDP book, 1986**

# Another Key Concept

- Structured embedding vectors via tensor-product rep.

symbolic semantic parse tree (complex relation)

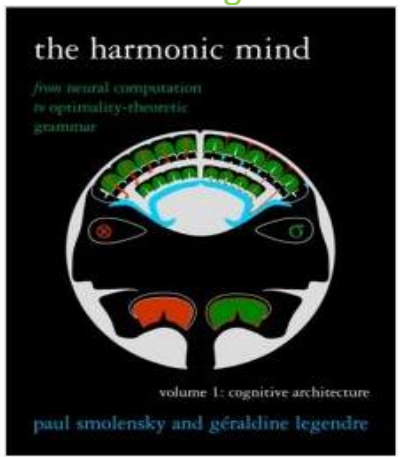
(e.g. binary trees w. theory)

Then, reasoning in symbolic-space (traditional AI) can be beautifully carried out in the continuous-space in human cognitive and neural-net (i.e., connectionist) terms

**Smolensky** & Legendre: **The Harmonic Mind, MIT Press, 2006**

From Neural Computation to Optimality-Theoretic Grammar

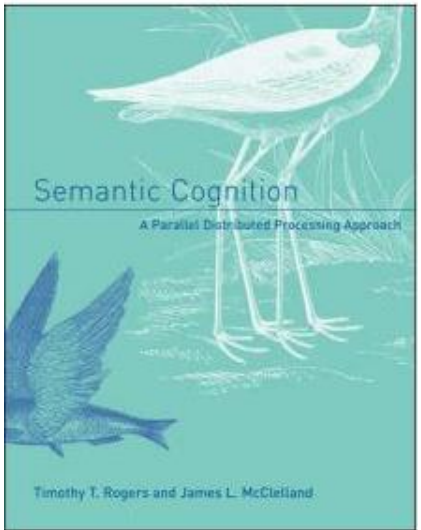
Volume I: Cognitive Architecture; Volume 2: Linguistic Implications




Rogers & McClelland

**Semantic Cognition**

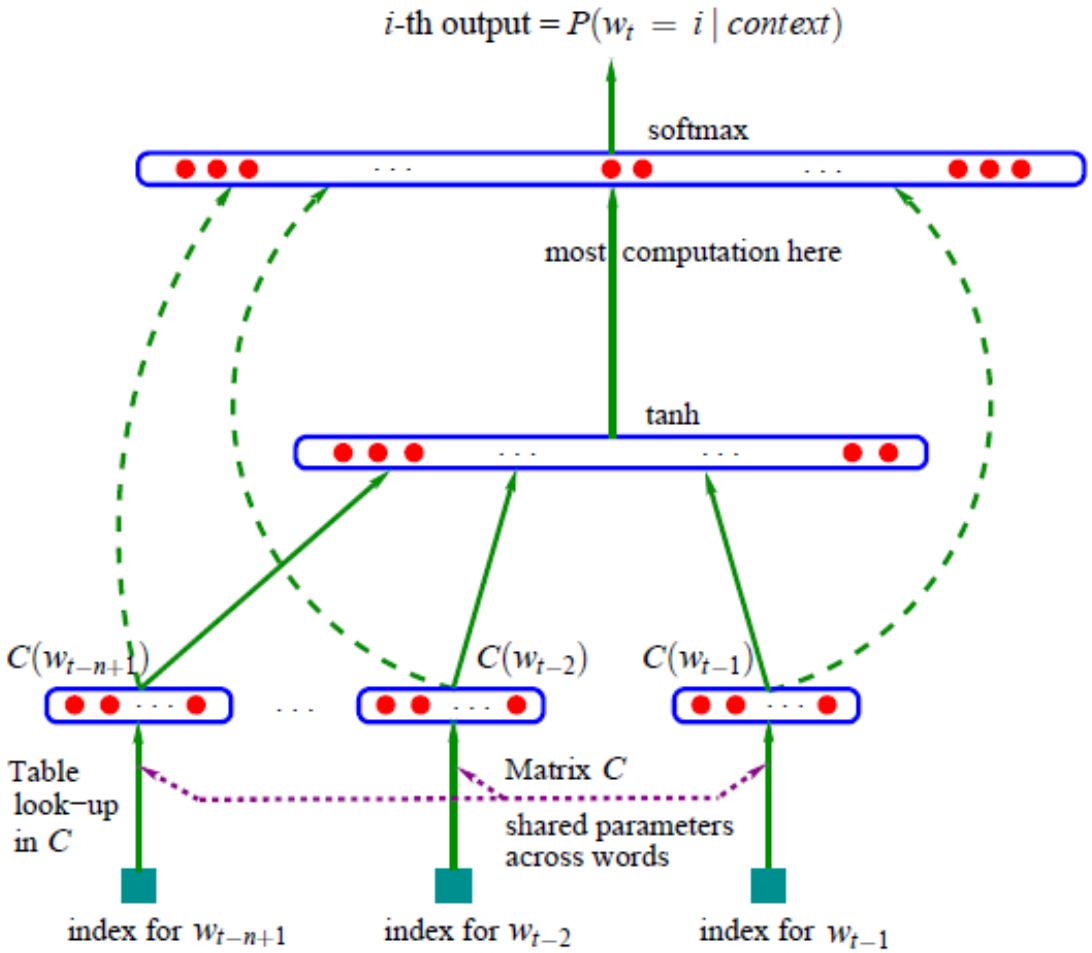
**MIT Press, 2006**



Journal of Machine Learning Research 3 (2003) 1137–1155 also NIPS-2001

# A Neural Probabilistic Language Model

**Yoshua Bengio** BENGIOY@IRO.UMONTREAL.CA
**Réjean Ducharme** DUCHARME@IRO.UMONTREAL.CA
**Pascal Vincent** VINCENTP@IRO.UMONTREAL.CA
**Christian Jauvin** JAUVINC@IRO.UMONTREAL.CA

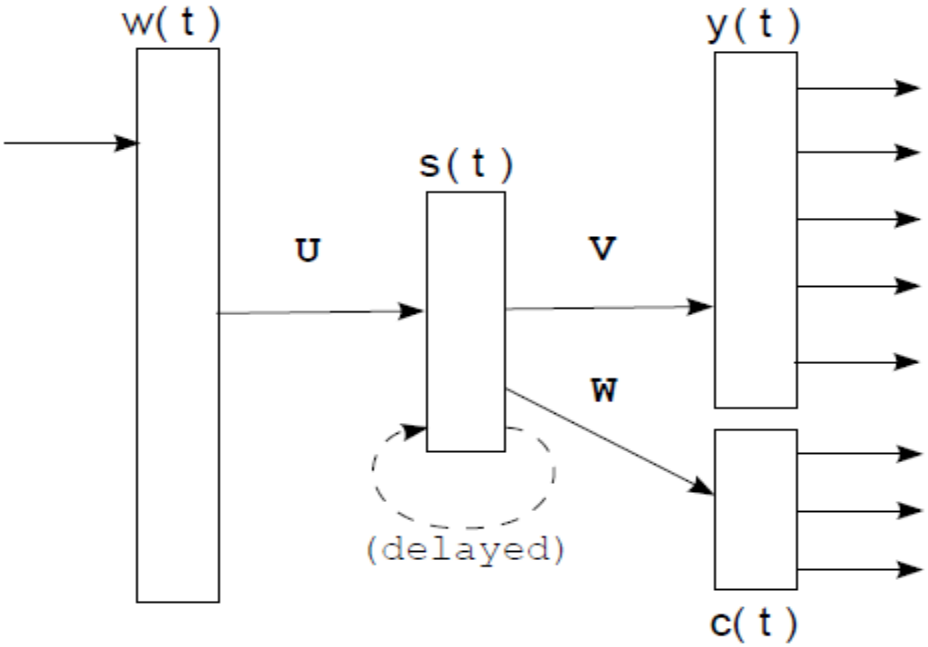



- Feed-forward NN
- Embedding: a by-product of n-gram prediction
- BP to learn embedding embedding vectors
- Output layer huge: |Vocab|
- Overcome by a later method (Collobert/Weston, 2008)

# Recurrent neural network based language model

*Tomáš Mikolov*[1,2], *Martin Karafiát*[1], *Lukáš Burget*[1], *Jan "Honza" Černocký*[1], *Sanjeev Khudanpur*[2]

[1]Speech@FIT, Brno University of Technology, Czech Republic
[2] Department of Electrical and Computer Engineering, Johns Hopkins University, USA




- RNN-LM also gives embedding
- RNN:FFNN=IIR:FIR filters

- Large LM perplexity reduction
- Lower ASR WER improvement
- Expensive in learning
- Later turned to FFNN at Google: Word2vec, Skip-gram, etc.
- All UNSUPERVISED

$$P(w_i|history) = P(c_i|\mathbf{s}(t))P(w_i|c_i,\mathbf{s}(t))$$

# Selected Research on Semantic Embedding at MSR

- DSSM (Deep-Structured Semantic Model)
- Deriving embedding vectors in a weakly supervised manner
- Smart exploitation of "supervision" signals (at virtually no cost)
- Performing much better than unsupervised embedding in Info retrieval (Bing web search), machine translation, and multimodal processing tasks

**References:**


Huang, He, Gao, Deng, Acero, Heck, Learning Deep Structured Semantic Models for Web Search using **Clickthrough Data**, ACM-CIKM, 2013

Shen, He, Gao, Deng, Mesnil, A Latent Semantic Model with Convolutional-Pooling Structure for Information Retrieval , ACM-CIKM, 2014

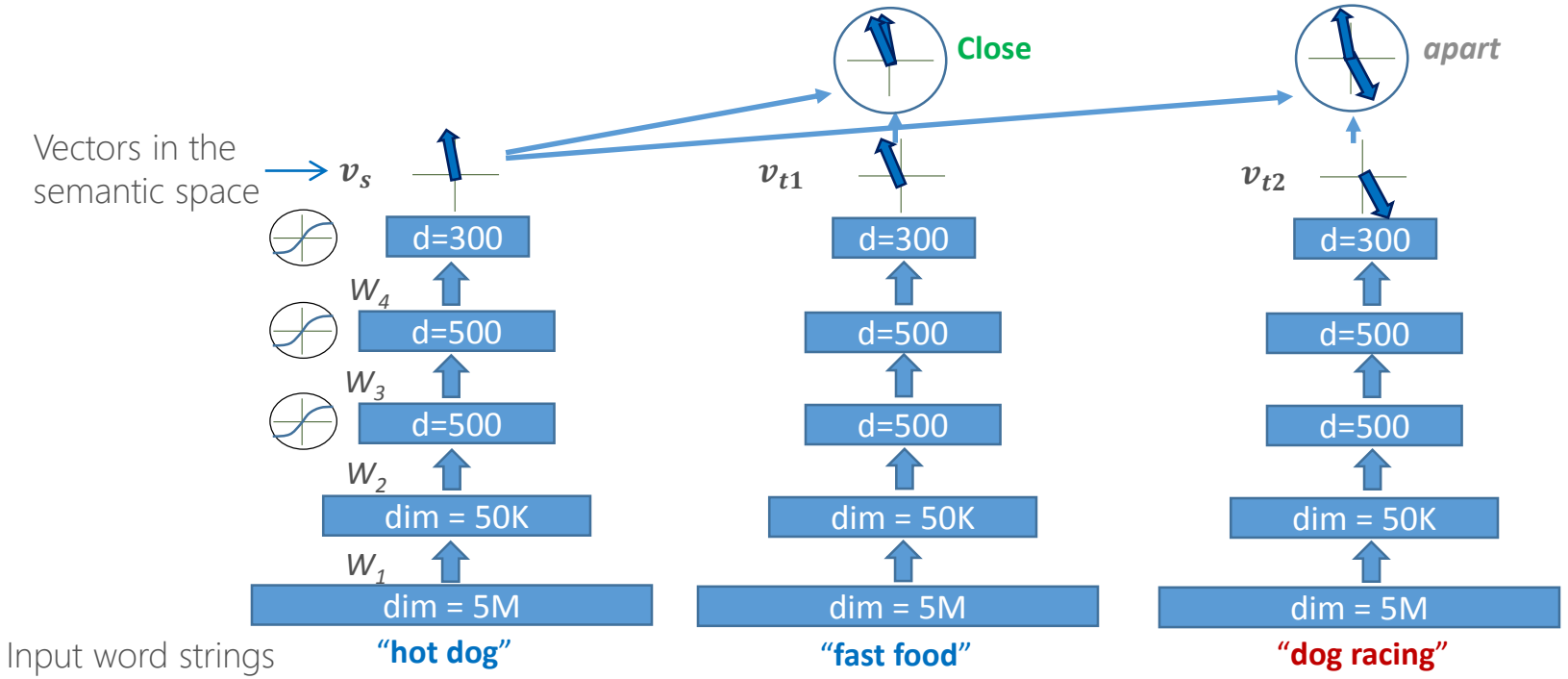
Gao, Pantel, Gamon, He, Deng, Shen, Modeling Interestingness with Deep Neural Networks, EMNLP, 2014

Gao, He, Yih, Deng, Learning Continuous Phrase Representations for **Translation** Modeling, ACL, 2014.

Shen, He, Gao, Deng, Mesnil, Learning Semantic Representations Using Convolutional Neural Networks for Web Search, WWW 2014.

# "hot dog" & "fast food": Close in semantic space

"hot dog" & "dog racing": *apart* (although sharing same word "dog")



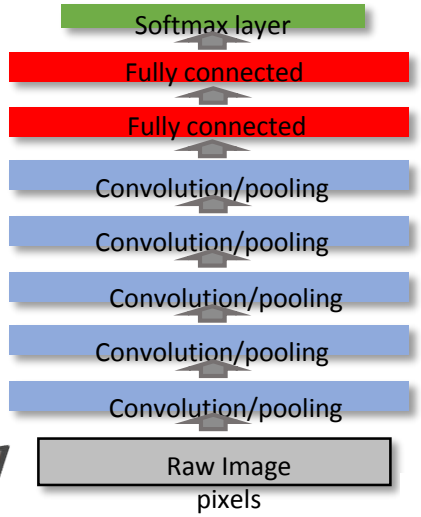


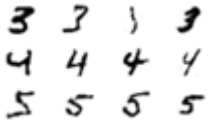

$$L(\Lambda) = -\log \prod_{(Q,D^+)} \frac{\exp[\psi R_\Lambda(Q, D^+)]}{\sum_{D' \in \boldsymbol{D}} \exp[\psi\, R_\Lambda(Q, D')]}$$

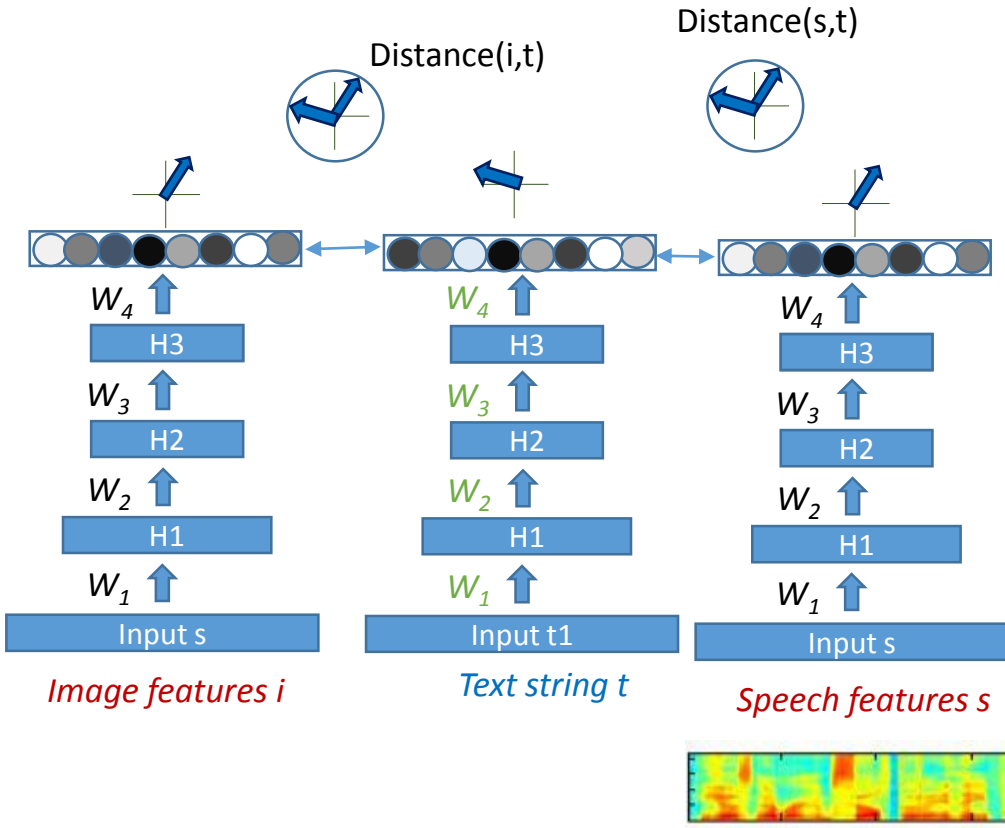
# DSSM for **Multi-Modal Learning (text, image, speech)**

## --- a "human-like" speech acquisition & image understanding model

- Recall DSSM for text inputs: *s*, $t_1$, $t_2$, $t_3$, …
- Now: replace text s by image *i* & speech *s*
- Using DNN/CNN features of image/speech



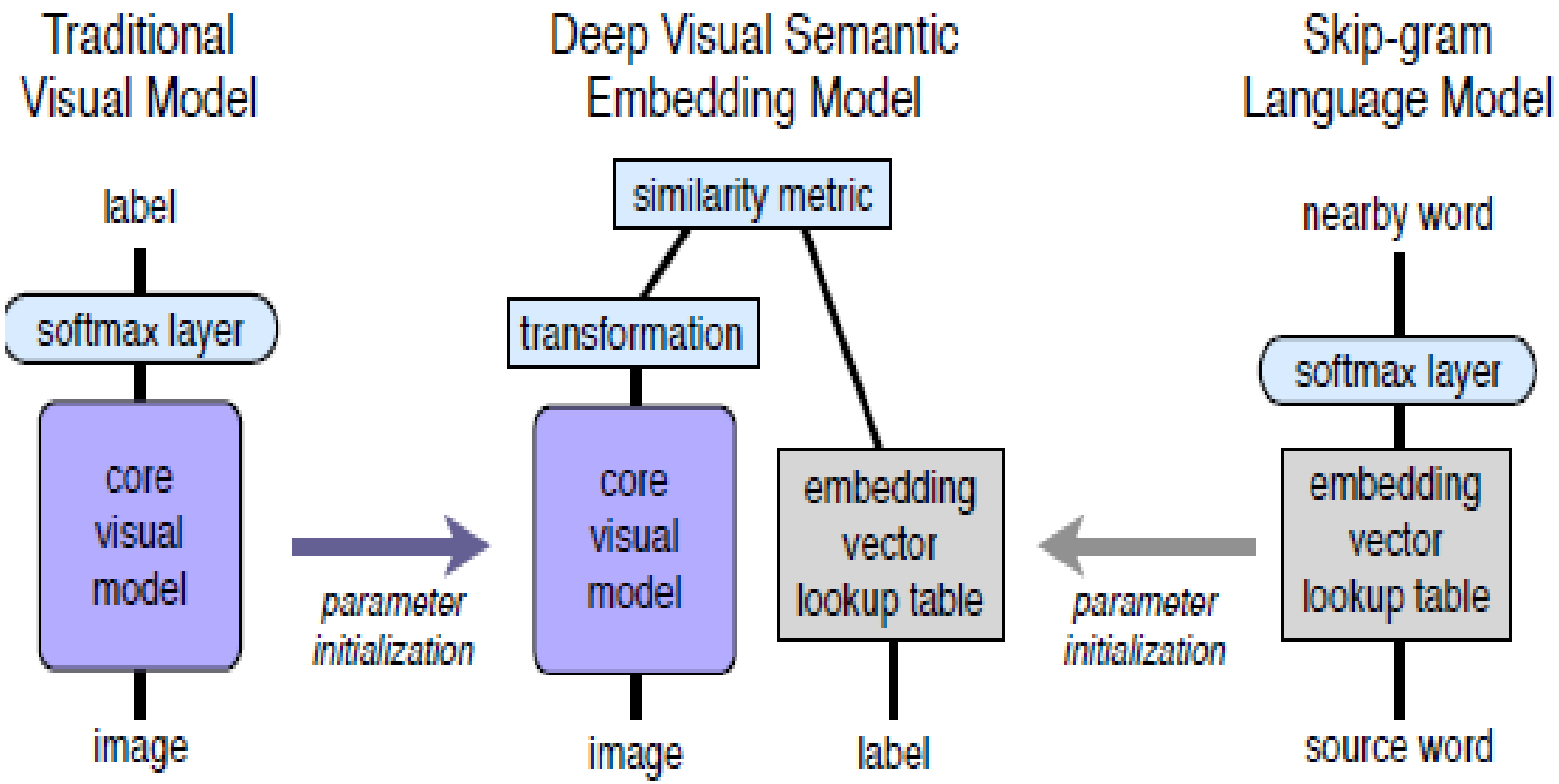


NIPS-2013

# DeViSE: A Deep Visual-Semantic Embedding Model

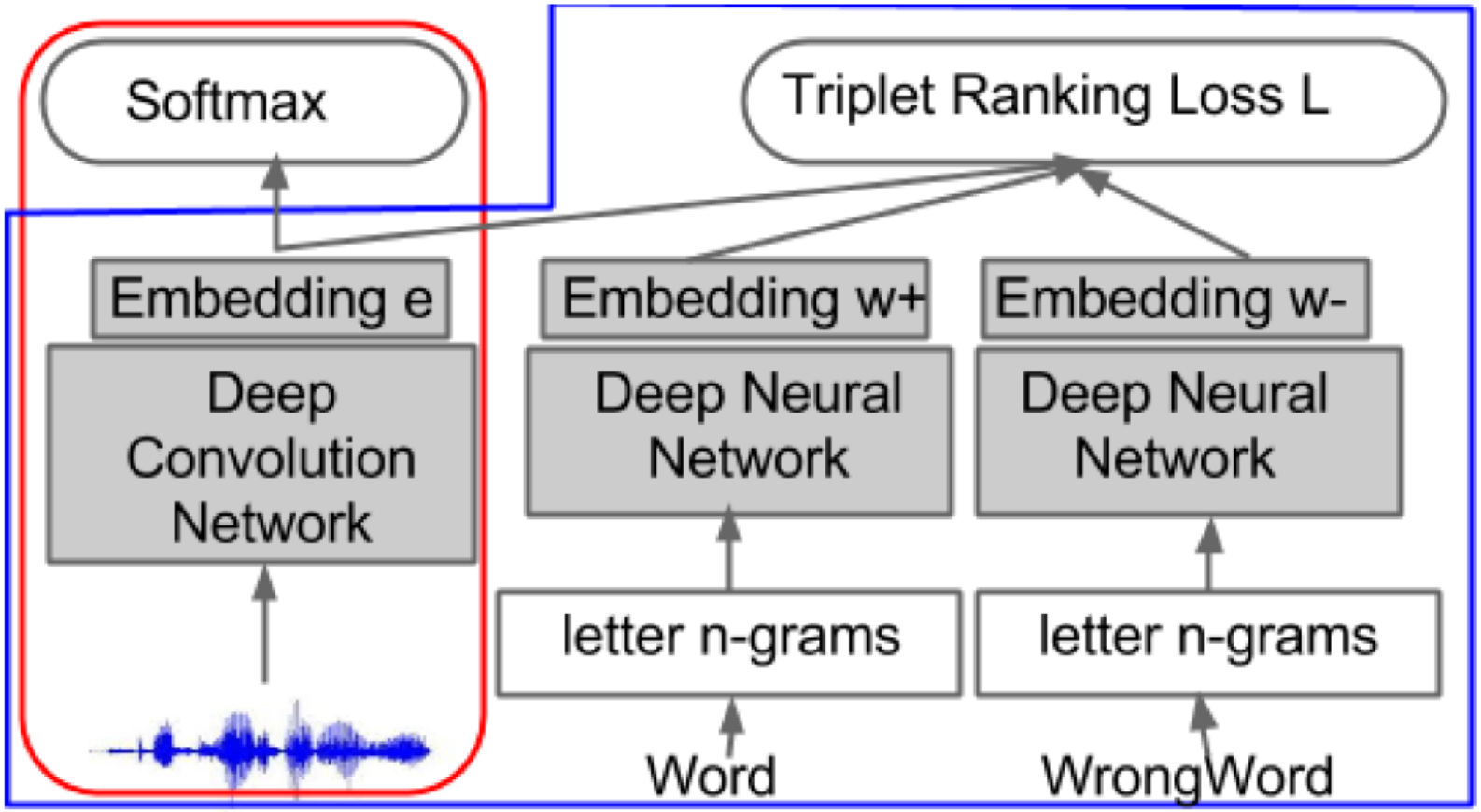
**Andrea Frome*, Greg S. Corrado*, Jonathon Shlens*, Samy Bengio**
**Jeffrey Dean, Marc'Aurelio Ranzato, Tomas Mikolov**
* These authors contributed equally.

{afrome, gcorrado, shlens, bengio, jeff, ranzato, tmikolov}@google.com
Google, Inc.
Mountain View, CA, USA



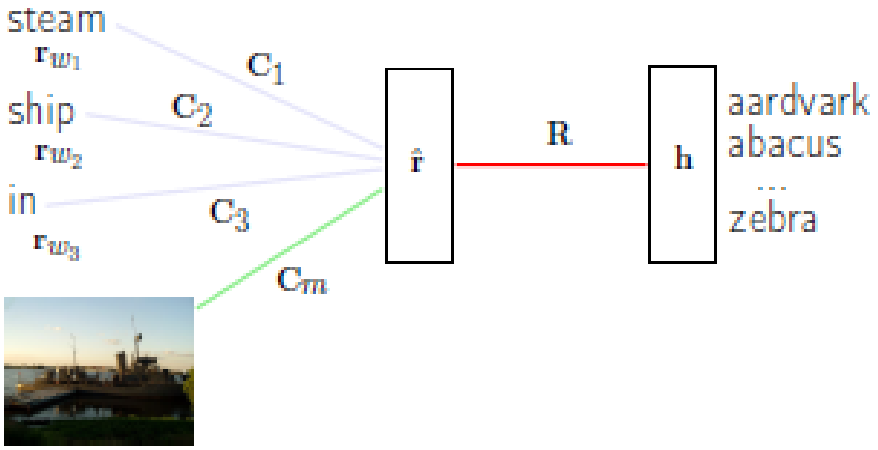
# Word Embeddings for Speech Recognition

*Samy Bengio and Georg Heigold*

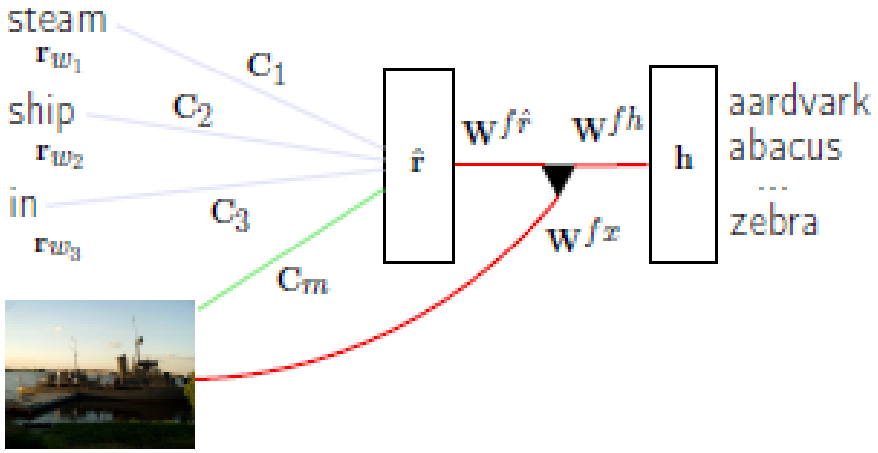


# Multimodal Neural Language Models

**Ryan Kiros** RKIROS@CS.TORONTO.EDU
**Ruslan Salakhutdinov** RSALAKHU@CS.TORONTO.EDU
**Richard Zemel** ZEMEL@CS.TORONTO.EDU
Department of Computer Science, University of Toronto
Canadian Institute for Advanced Research




(a) Modality-Biased Log-Bilinear Model (MLBL-B)




(b) Factored 3-way Log-Bilinear Model (MLBL-F)

*Figure 2.* Our proposed models. Left: The predicted next word representation $\hat{\mathbf{r}}$ is a linear prediction of word features $\mathbf{r}_{w_1}, \mathbf{r}_{w_2}, \mathbf{r}_{w_3}$ (blue connections) biased by image features $\mathbf{x}$. Right: The word representation matrix $\mathbf{R}$ is replaced by a factored tensor for which the hidden-to-output connections are gated by $\mathbf{x}$.

# Many possible applications of DSSM: Learning semantic similarity between *X* and *Y*

| Tasks | *X* | *Y* |
|---|---|---|
| **Web search** | ***Search query*** | ***Web documents*** |
| **Ad selection** | ***Search query*** | ***Ad keywords*** |
| **Entity ranking** | ***Mention (highlighted)*** | ***Entities*** |
| **Recommendation** | ***Doc in reading*** | ***Interesting things in doc or other docs*** |
| **Machine translation** | ***Sentence in language A*** | ***Translations in language B*** |
| Nature User Interface | *Command (text/speech)* | *Action* |
| Summarization | *Document* | *Summary* |
| Query rewriting | *Query* | *Rewrite* |
| Image retrieval | *Text string* | *Images* |
| … | … | … |

# Functional Modeling of the Brain/Mind

- For deep intelligence: reasoning, beyond classification and similarity measure/ranking
- Reasoning at symbolic level (reasonably well understood by AI)
- But how does the brain use neural nets to do symbolic computation?
- Three levels of brain-function abstractions:
  - Neural-activity level (e.g., DNN)
  - Vector level (DSSM for entity/concept & multimodal information embedding)
  - Symbolic level (tensor product representation)
- From strongly supervised learning (speech/vision problems)

  to weakly supervised learning (language and multimodal problems)

  to unsupervised learning (reasoning and AI problems)

# A Big Picture

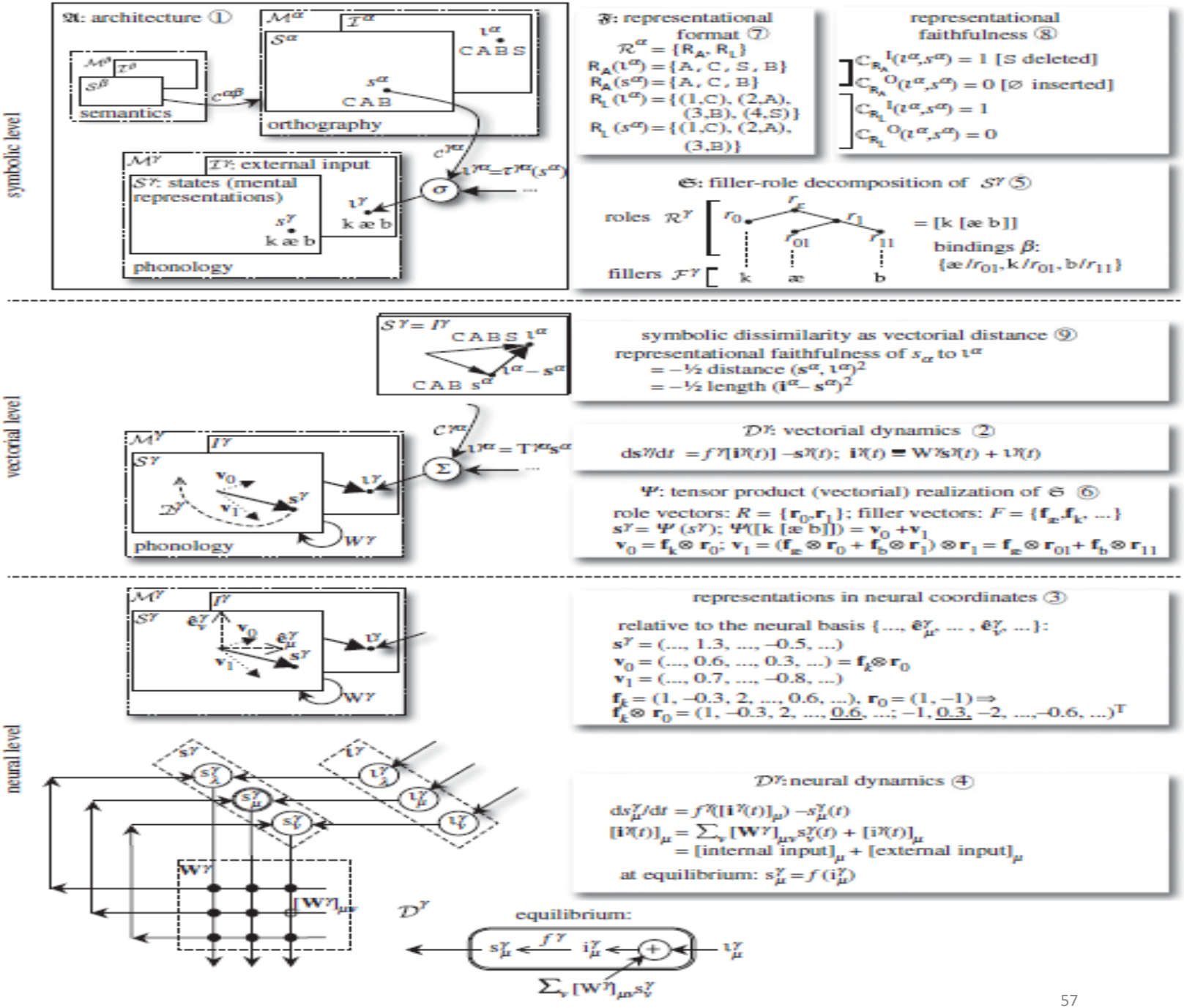



Figure 1. A schematic of the proposed theory for neural computation of symbolic

Smolensky et al., 1986-2013

# Main message of this talk

## Deep Learning

## =

$$\mathcal{F}[\ldots\mathcal{F}[\text{Deep-Neural-Net};\text{Deep-Generative-Model}]\ldots]$$

| | | |
|---|---|---|
| **Speech recognition** | **RNN, LSTM** | **HDM (w. new formulation & paramet.)** |
| **Speech enhancement** | **DNN/DSN (feedforw'd)** | **Unfolded non-negative matrix factorization** |
| **Language/multimodal** | **DSSM** | **(Hierarchical) topic models** |
| **Algorithms** | **BackProp,…** | **BP & BP (BeliefProp, multiplicative units), stochastic variational EM (BackP in E-step)** |
| **Neuroscience** | **"Wake"** | **"Sleep"** |

**Two seemingly opposing approaches: with highly complementary strengths and weaknesses**

# Thank You

## Q/A & discussions